

# 蛋白質結構比對搜尋

## Protein Structure Comparison and Search

呂平江

國立清華大學

生命科學系/生物資訊與結構生物研究所

2013/06/26

# Sequence Comparison

```

*****
1ATP_E_House__Mus  GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1APH_E_House__Mus  GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
2CPK_E_House__Mus  GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1JLU_E_House__Mus  GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1JBP_E_House__Mus  GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1YDR_E_Cow__Bos    GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDFERIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1YDS_E_Cow__Bos    GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDFERIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1YDT_E_Cow__Bos    GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDFERIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1CDK_A_Pig__Sus     GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDFERIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80
1CDK_B_Pig__Sus     GMAAAAKKGSSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDFERIKTLGTGSFGRVHLVKHKEGNNHYANKILDQKQVV 80

```



```

*****
1ATP_E_House__Mus  KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1APH_E_House__Mus  KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
2CPK_E_House__Mus  KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1JLU_E_House__Mus  KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1JBP_E_House__Mus  KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1YDR_E_Cow__Bos    KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1YDS_E_Cow__Bos    KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1YDT_E_Cow__Bos    KLKQIEHTLNKKRILQAVNFPFLVKLEFSEKDNNSMLYHVMHEYVAGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1CDK_A_Pig__Sus     KLKQIEHTLNKKRILQAVNFPFLVKLEYSEKDNNSMLYHVMHEYVPGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160
1CDK_B_Pig__Sus     KLKQIEHTLNKKRILQAVNFPFLVKLEYSEKDNNSMLYHVMHEYVPGGEMFSLRRIIGRFSEPHARFYAAQIVLTFFEYLNHSL 160

```



```

*****
1ATP_E_House__Mus  DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWLTCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1APH_E_House__Mus  DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWLTCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
2CPK_E_House__Mus  DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWLTCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1JLU_E_House__Mus  DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1JBP_E_House__Mus  DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1YDR_E_Cow__Bos    DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1YDS_E_Cow__Bos    DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1YDT_E_Cow__Bos    DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1CDK_A_Pig__Sus     DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWLTCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240
1CDK_B_Pig__Sus     DLIYRDLKPEMLLIDQQGYIQVTDGFAKRVKGRTWLTCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAAGYPPFFFA 240

```



% of Identity

% of similarity

# Introduction to Structure Comparison

- Sander & Schneider (1990) :

Total available structures : 597

Protein sequences (PDB) → Pairwise alignment

sequence identity > 30% 100% have similar 3D-structure

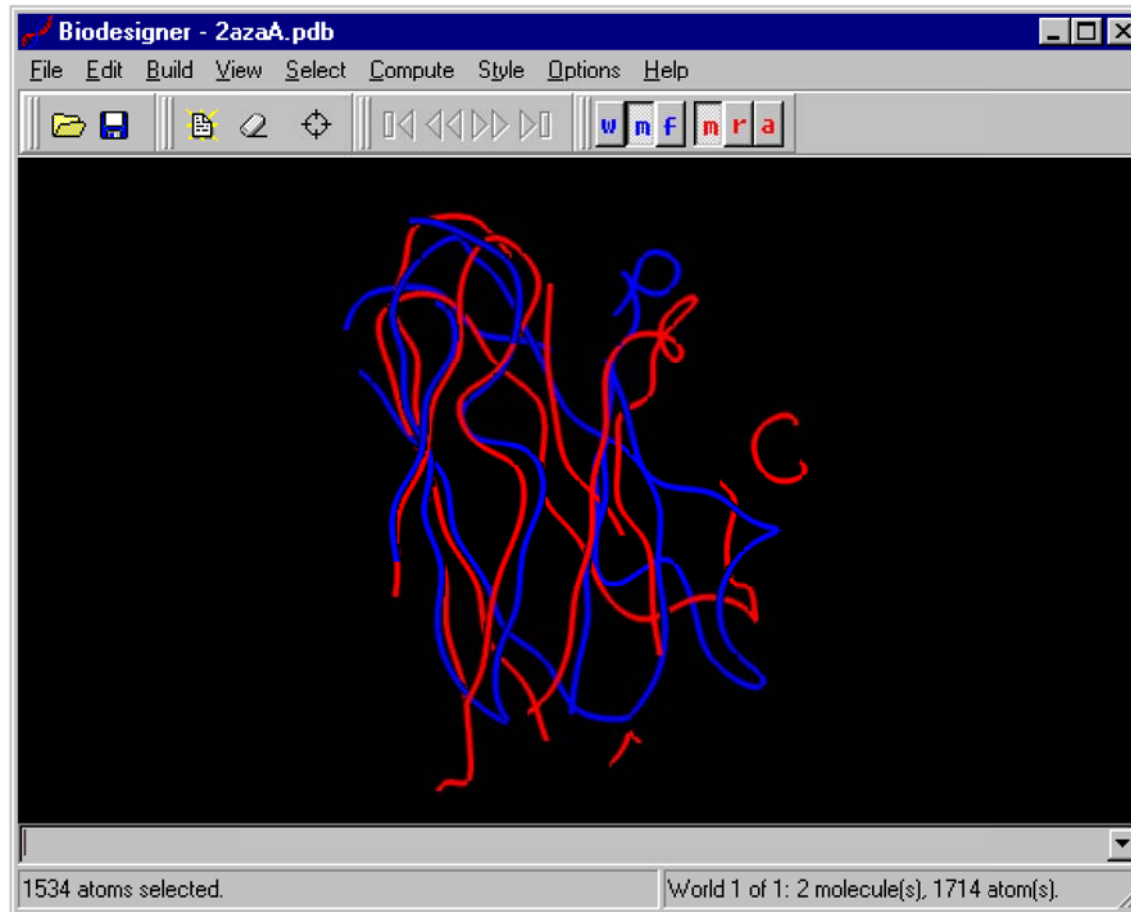
- Burkhard Rost (1999) :

Total available structures : 11,364

sequence identity > 30% 90% have similar 3D-structure

sequence identity < 25% 10% have similar 3D-structure

# Structure Comparison



蛋白質結構比對是用來比較兩個蛋白質結構是否相似，通常是計算兩結構距離的方均根差(RMSD, Root Mean Square Deviation)，而其單位通常是埃( $\text{\AA}$ )。



# Structural Comparisons – Why?

- Finding similar structure by sequence comparison alone is not enough
- The number of protein structures is increasing rapidly.
- Structure-Structure relationship is more conserved than sequence during evolution

# Structural Comparisons – How?

- Two categories of current methods
  - By amino acid sequence alignments.
  - By 3D structural alignments.

# Classical Sequence Alignment Methods

- BLAST
  - Basic Local Alignment Search Tool
- FASTA
  - FAST-All, reflecting that it can be used for fast protein comparisons

**Performance: Rapid but inaccurate\***

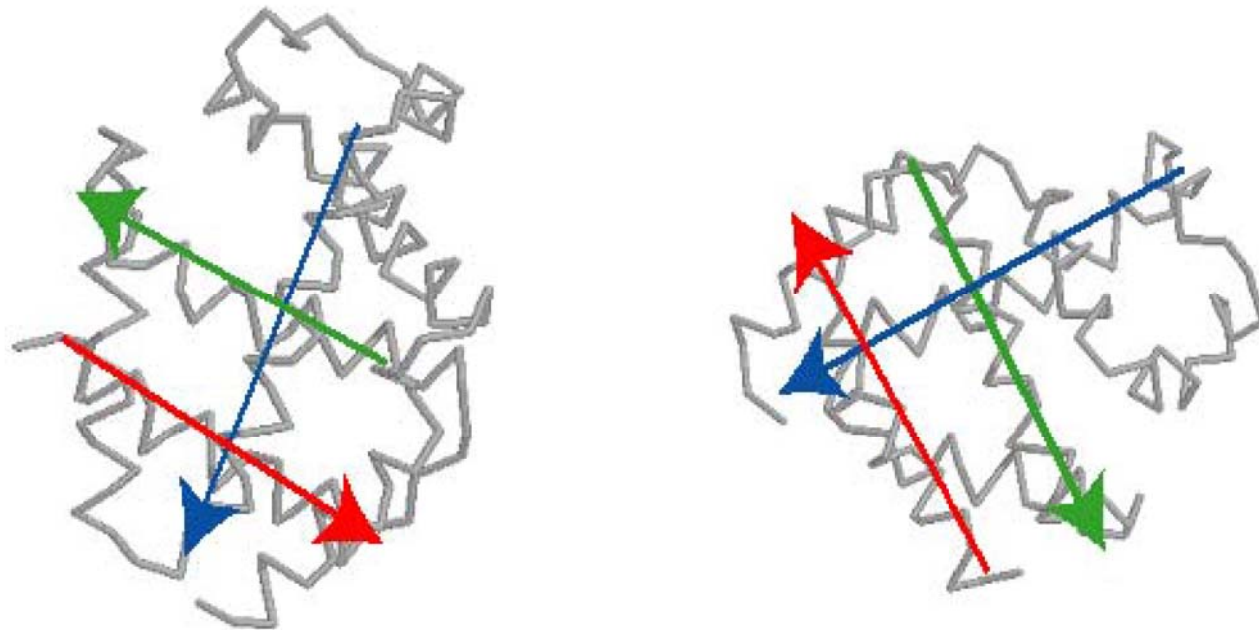
\* Kolodny *et al.* (2005) *J Mol Biol.* 346:1173-1188

# Conventional Structural Alignment Methods

- Double Dynamic Programming – SSAP
- Distance Alignment Tools – DALI
- Vector Alignment Search Tool – VAST
- Combinatorial Extension – CE
- Fast Alignment Search Tool – FAST
- MAtching Molecular Models Obtained from Theory – MAMMOTH

# Structure Comparison Methods (VAST)

- VAST (Vector Alignment Search Tool ) (Gibrat, Madej,1996)
- Secondary Structure Elements (SSE)



**Vector Alignment Search Tool**

<http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>

## Structures similar to VAST Search VS29688 , VS29

View / Save Alignments

[new Get Cn3D 4.0!](#)

### Options: (1)

- ☒ Launch Viewer
- ☐ See File
- ☐ Save File

### Viewer: (2)

- ☒ Cn3D (asn.1)
- ☐ Mage (Kinemage)
- ☐ (PDB)

### Complexity: (3)

- ☒ Aligned Chains only
- ☐ All Chains
- ☒ Alpha Carbons only
- ☐ All Atoms

Structure neighbors 1-5 out of 5 displayed. Page 1 of 1.

(4)

| <input type="checkbox"/> | PDB                  | C | D | RMSD | NRES | %Id   | Description                                                                                                                                                 |
|--------------------------|----------------------|---|---|------|------|-------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/> | <a href="#">1A1X</a> |   |   | 0.0  | 106  | 100.0 | Crystal Structure Of Mtcp-1 Involved In T Cell Malignancies                                                                                                 |
| <input type="checkbox"/> | <a href="#">1E8W</a> | A | 1 | 2.7  | 37   | 10.8  | Crystal Structure Of The Human 8-Oxoguanine Glycosylase (Hogg1) Bound To A Substr                                                                           |
| <input type="checkbox"/> | <a href="#">1RTH</a> | A | 2 | 1.1  | 20   | 0.0   | Hiv-1 Reverse Transcriptase Mol_id: 1; Molecule: Hiv-1 Reverse Transcriptase; Chain: A, B; Synonym: Hiv-1 Rt; Ec: 2.7.7.49; Engineered: Yes                 |
| <input type="checkbox"/> | <a href="#">1TF1</a> |   |   | 1.5  | 29   | 3.4   | Transcriptional Elongation Factor Sii (Tfiis, Nucleic-Acid Binding Domain) (Nmr, 12 Structures)                                                             |
| <input type="checkbox"/> | <a href="#">1BC3</a> | C |   | 1.2  | 23   | 4.3   | Structures Of Sap-1 Bound To Dna Sequences From The E74 And C-Fos Promoters Provide Insights Into How Ets Proteins Discriminate Between Related Dna Targets |

Display / Sort Hits

page number:  Hits to display per page:  choose between 20-100 neighbors per page.

### Display Subset: (5)

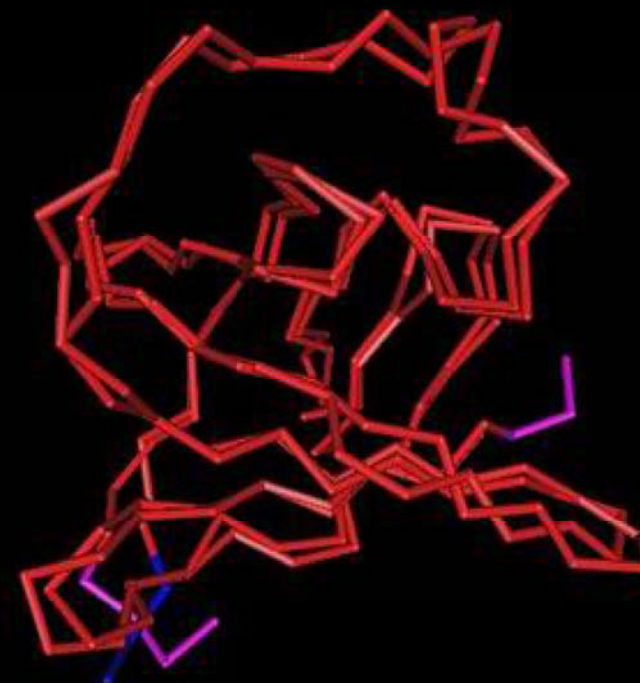
- ☒ Non-redundant; BLAST p-value 10e-7
- ☐ Non-redundant; BLAST p-value 10e-40
- ☐ Non-redundant; BLAST p-value 10e-80
- ☐ Non-identical sequences
- ☐ All of MMDB

### Sorted by: (6)

- ☒ VAST Score
- ☐ VAST P-value
- ☐ Rmsd
- ☐ Aligned residues
- ☐ Identities

### Column Format: (7)

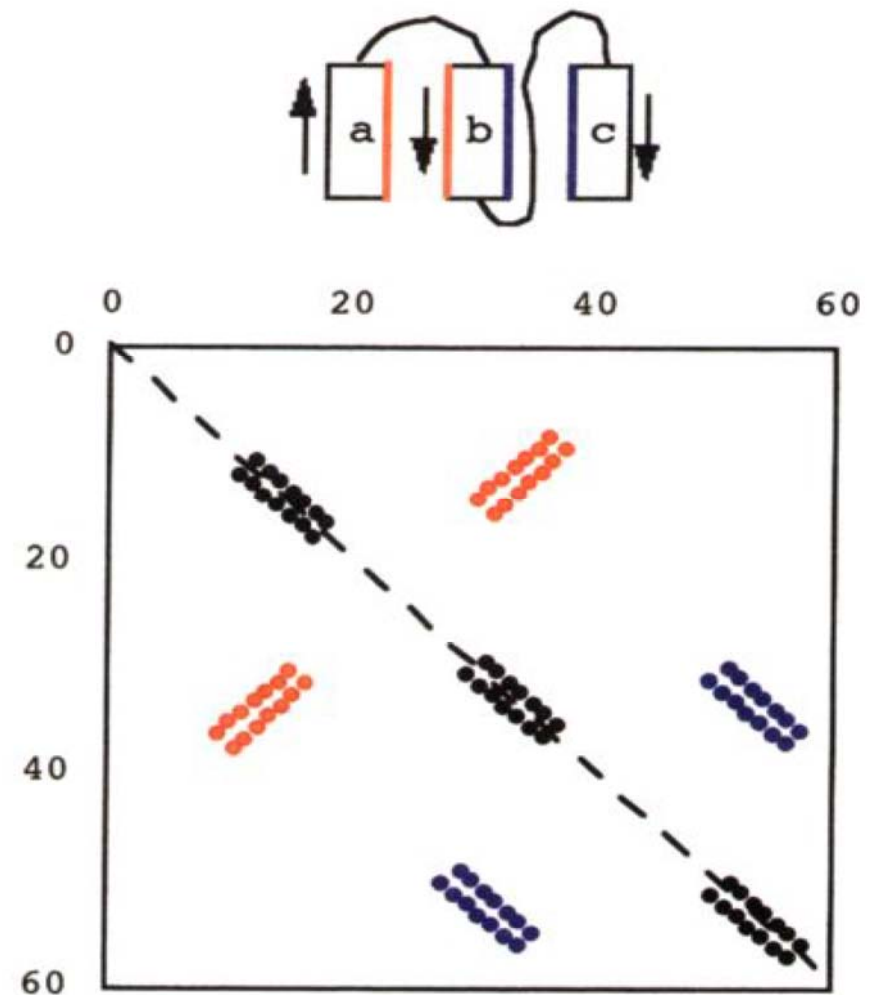
- ☒ RMSD, NRES, %Id
- ☐ All values





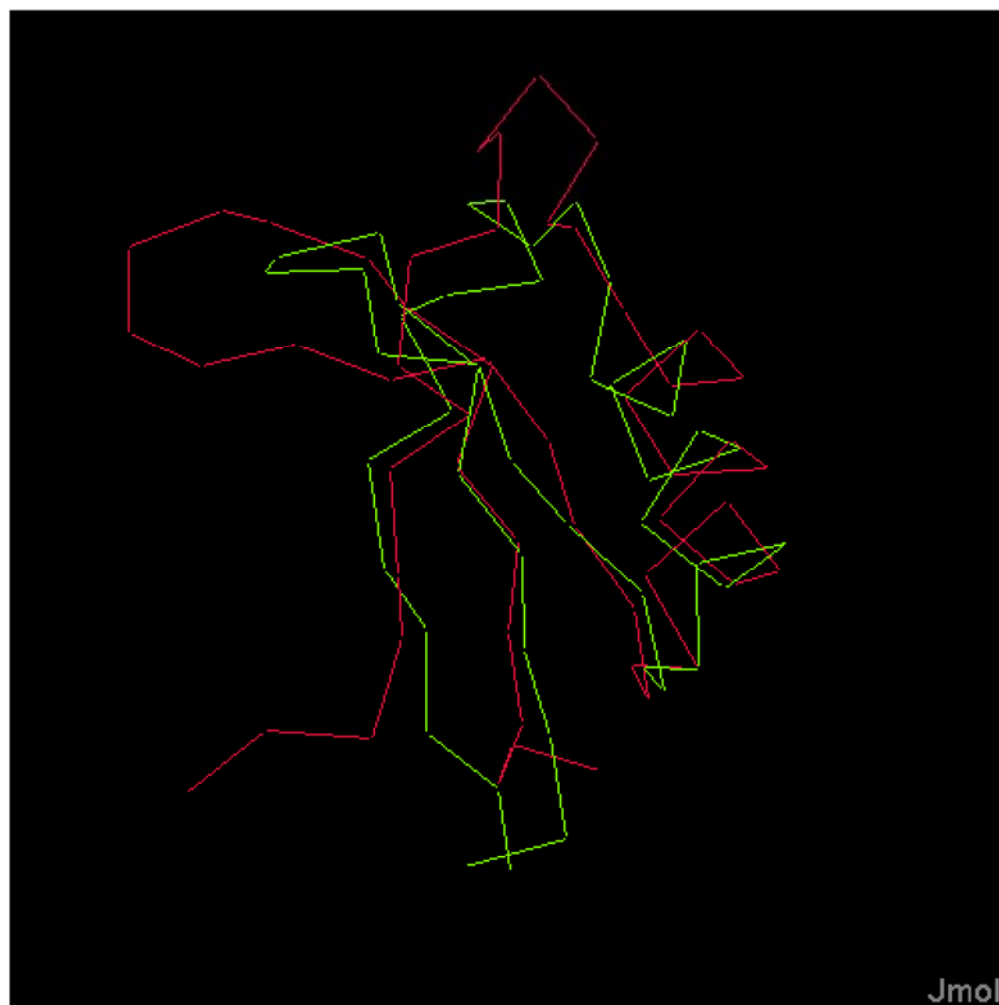
# Structure Comparison Methods (DALI)

- **DALI** (distance alignment tools)
  - Transfer all the C $\alpha$  - C $\alpha$  distance in a protein to **distance dot matrix**
  - If C $\alpha$ -C $\alpha$ distance > 12 Å  
`delete the dot`
  - 3D→2D



## DaliLite Results: Superimposed structures

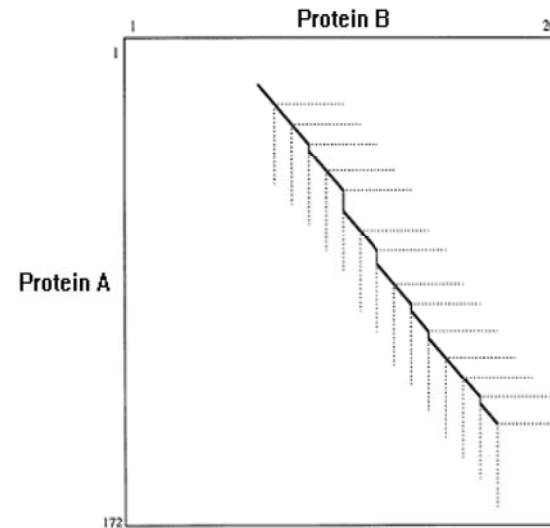
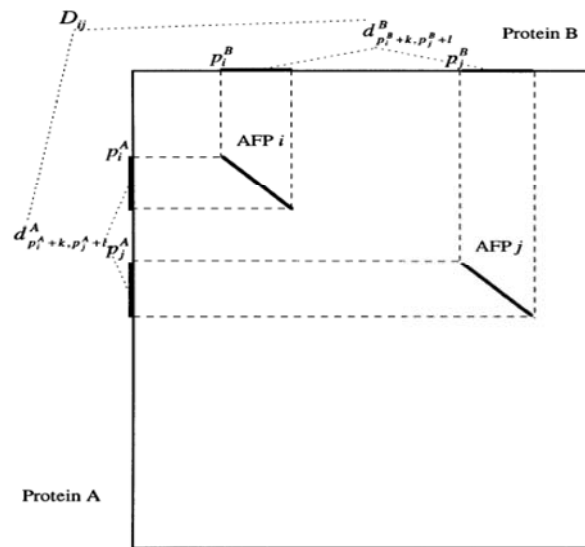
Starting a Jmol applet; it may take a few seconds. If you are loading many structures, you can monitor freezes due to running out of memory (see About Jmol -> Java memory usage), then close all Jmol again, or (ii) select fewer structures.



Toggle: ☐ spinning ☐ superimpose all ligands ☐ Clear labels

# Structure Comparison Methods (CE)

- combinatorial extension 
$$D_{ij} = \frac{1}{m^2} \left( \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| d_{p_i^A+k, p_j^A+l}^A - d_{p_i^B+k, p_j^B+l}^B \right| \right)$$



CE是一種計算結構比對的方法，是由美國聖地牙哥超級電腦中心所提供的。CE是利用區域的幾何特性(alpha碳原子間的向量)來進行比對，將配對到的片段稱為AFPs (aligned fragment pairs)，再利用演算法來得到最好的RMSD值。

# Protein structure classification databases

- **SCOP** (structure classification of proteins)
  - ✓ based on expert definition of structural similarities
  - ✓ <http://scop.mrc-lmb.cam.ac.uk/scop/>
- **CATH** (classification by class, architecture, topology and homology)
  - ✓ based on **SSAP** (secondary structure alignment program)
  - ✓ University College, London (Orengo 1997)
  - ✓ <http://www.biochem.ucl.ac.uk/bsm/cath/>
- **FSSP** (fold classification based on structure-structure alignment of proteins)
  - ✓ based on pairwise structure alignment of PDB by **DALI**
  - ✓ <http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html>
- **MMDB** (molecular modelling database)
  - ✓ PDB has been categorized into structurally related groups in MMDB by **VAST** (vector alignment search tool)
  - ✓ <http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml>
- **CE** (combinatorial extension)
  - ✓ based on pairwise structure alignment of PDB by **CE**
  - ✓ <http://cl.sdsc.edu/ce.html>

Searching similar protein structures of the  
specified protein in PDB

SARST – Structural similarity search  
Aided by  
Ramachandran Sequential Transformation

# Introduction to SARST

- SARST transforms 3D protein structures into 1D text sequences and recruit blast to perform protein structural alignment searches
- Features
  - high speed
  - reasonable compromise of the accuracy
  - giving statistically meaningful results



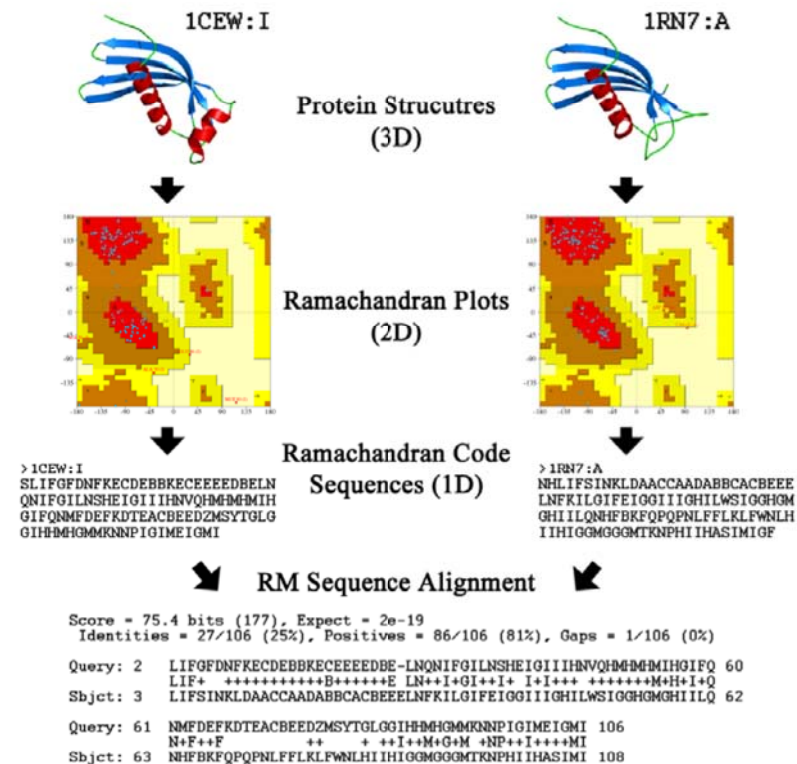
# Speed vs. Accuracy: Incompatible?

- Possible solution: the linear encoding method  
**3D structure → 1D text sequence**

- Example:

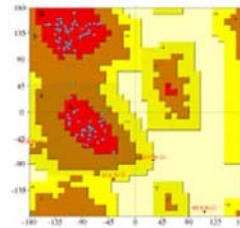
## SARST

- Structure Alignment by  
Ramachandran  
Search Tool





Query Protein  
Strucutre (3D)



Ramachandran Plot  
(2D)

# SARST

Pre-transformed Structure Database

## • SARST

- Structural similarity search Aided by

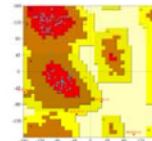
Ramachandran  
Sequential  
Transformation

```
>Query Protein
SLIFGDFNFKCEDBBKEEEDBELN
QNI FGI LNSHEIGIIINVOQHMMH
GIFQNMDFDKTEACBEEDZMSYGLG
GIHMHMMKNNPISIMEIGMI
Ramachandran Code
Sequence (1D)
```

Database Se



Query Protein  
Strucutre (3D)



Ramachandran Plot  
(2D)

```
>Query Protein
SLIFGDFNFKCEDBBKEEEDBELN
QNI FGI LNSHEIGIIINVOQHMMH
GIFQNMDFDKTEACBEEDZMSYGLG
GIHMHMMKNNPISIMEIGMI
Ramachandran Code
Sequence (1D)
```

Database Search

Pre-transformed Structure Database

```
101M:_ GHQCBABBACDEBBDECKEMAAABACCABCCBEEMDD...
117E:A HIGMHGYHFPPIEKLIIMHMPPHFILKENLFAHEMCK...
121P:_ FGHHHGHYFLPKWCABCAACCCENPILLMIHMFEMIG...
13PK:A FFIDDKLHNLPHLHLMHIMNNFLOGPPGLKMNDEEDK...
15C8:L LHHHHQEGIGIFFWFLGIIIMGHKIIILDRNHIGGGH...
16VP:A KMFFFFLLIFDEEACCDADABKPNDRKCCACBBDEKIN...
19HC:A FFFFFBWFFAIGLHIMLSNFFPMNPLSFFFFHBLCA...
1A00:A LQBCACCCACABEAERKKCDACBACAACCBEMDD...
1B44:D FNIBABBDKKIHPQSLMLEPSFFEGHGGGNFLPMQGG...
1CEW:I SLIFGDFNFKCEDBBKEEEDBELNQNI FGI LNSH...
1D06:A HDCBABAABAARNFDBBEDKKLCFFGGGIGKWLNL...
1ESU:I HMGLENMHHFDCNPHMCHLNIHMYCZGIGIHGGZVEV...
```

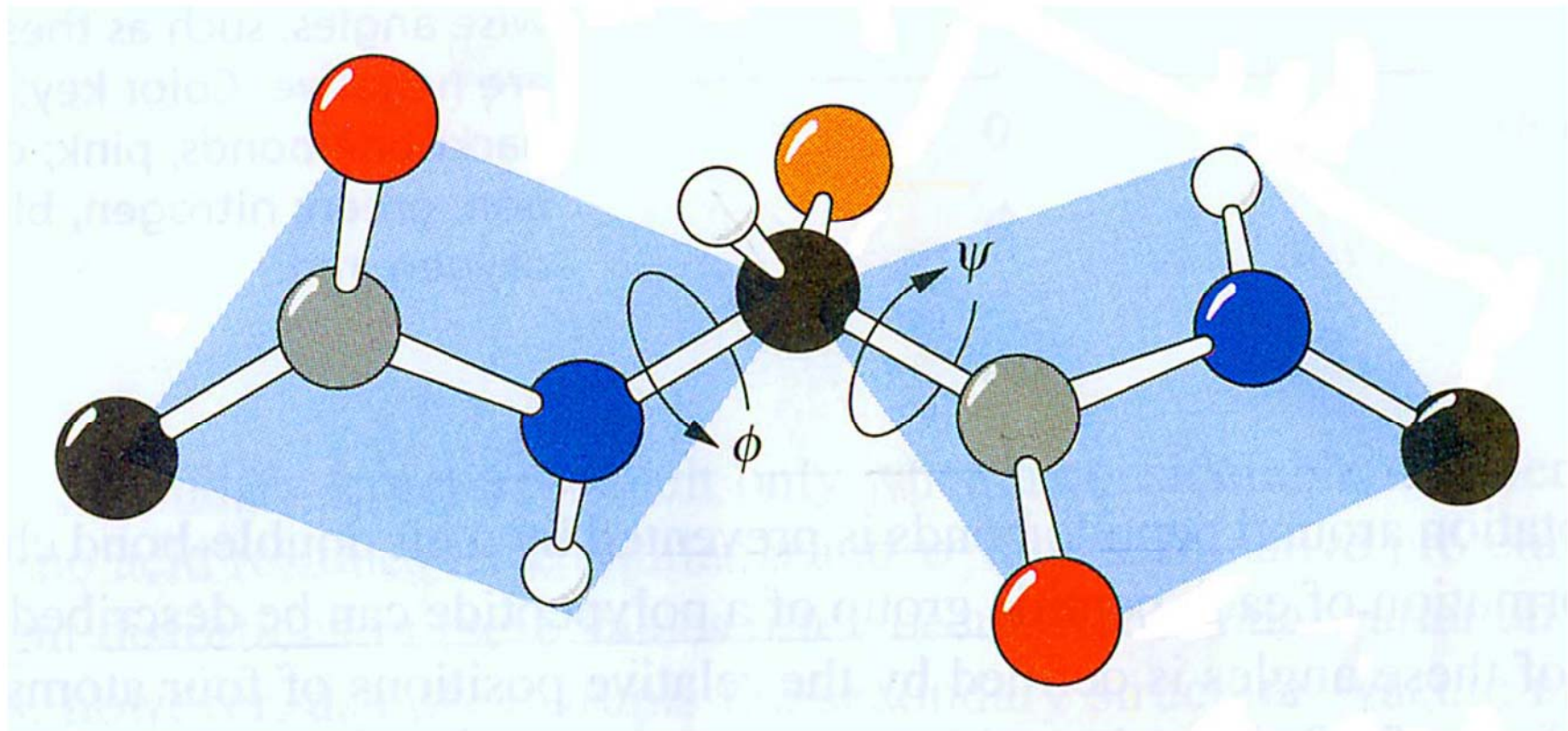
Wei-Cheng Lo (羅佳正), 2006

| No | PDB ID | Chain | Score | E-value |
|----|--------|-------|-------|---------|
| 1. | 1CEW   | I     | 180   | 3e-46   |
| 2. | 1R4C   | E     | 85    | 1e-17   |
| 3. | 2CH9   | A     | 84    | 2e-17   |
| 4. | 1R4C   | H     | 82    | 1e-16   |
| 5. | 1YVB   | I     | 81    | 2e-16   |

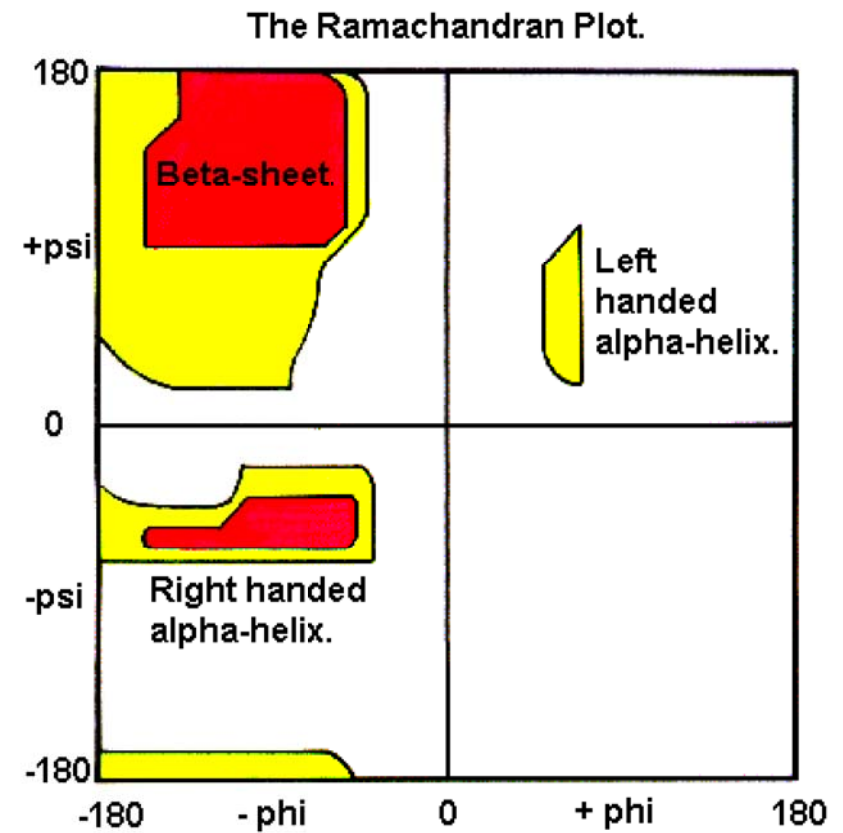
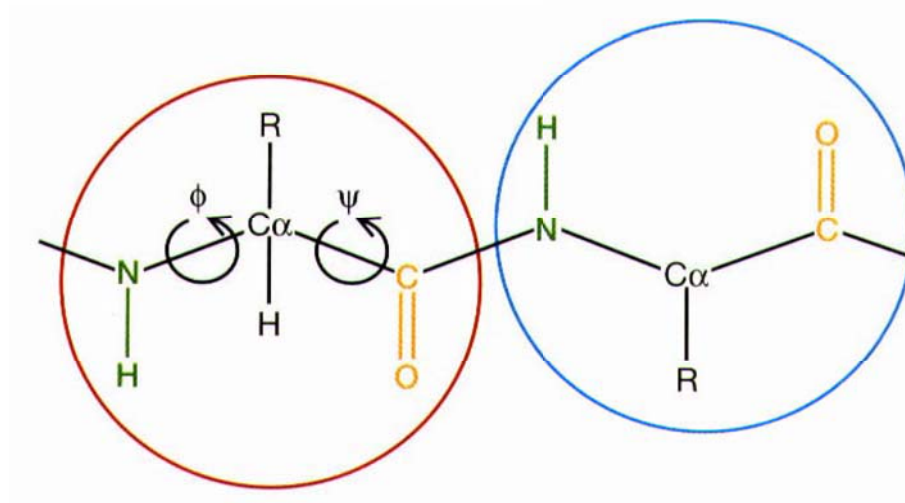
| No. | PDB ID | Chain | Score | E-value | Description                     | Organism      |
|-----|--------|-------|-------|---------|---------------------------------|---------------|
| 1.  | 1CEW   | I     | 180   | 3e-46   | Cystatin (Proteinase Inhibitor) | Gallus gallus |
| 2.  | 1R4C   | E     | 85    | 1e-17   | Cystatin C with Domain Swapping | Homo sapiens  |
| 3.  | 2CH9   | A     | 84    | 2e-17   | Cystatin F                      | Homo sapiens  |
| 4.  | 1R4C   | H     | 82    | 1e-16   | Cystatin C with Domain Swapping | Homo sapiens  |
| 5.  | 1YVB   | I     | 81    | 2e-16   | Cystatin                        | Gallus gallus |

|                                 |               |
|---------------------------------|---------------|
| Cystatin (Proteinase Inhibitor) | Gallus gallus |
| Cystatin C with Domain Swapping | Homo sapiens  |
| Cystatin F                      | Homo sapiens  |
| Cystatin C with Domain Swapping | Homo sapiens  |
| Cystatin                        | Gallus gallus |

# Phi ( $\phi$ ) and Psi ( $\psi$ ) Angles

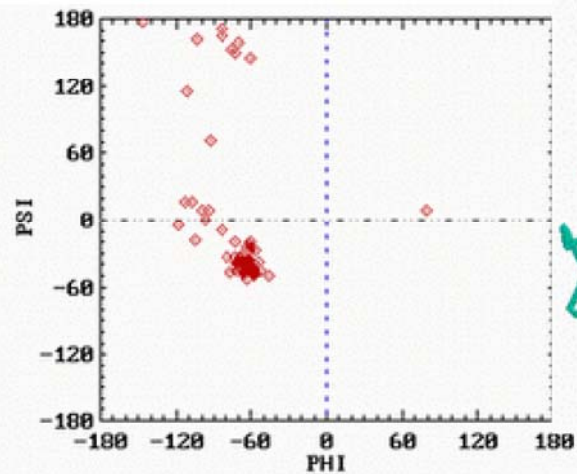


# Ramachandran Plot

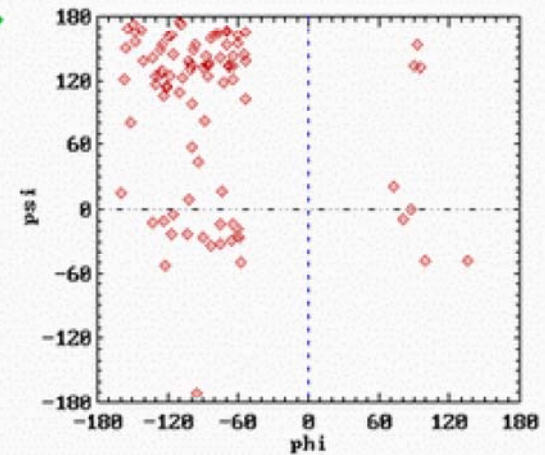
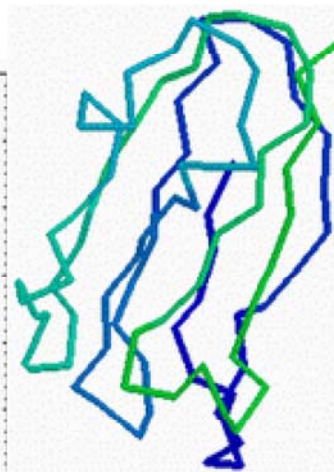




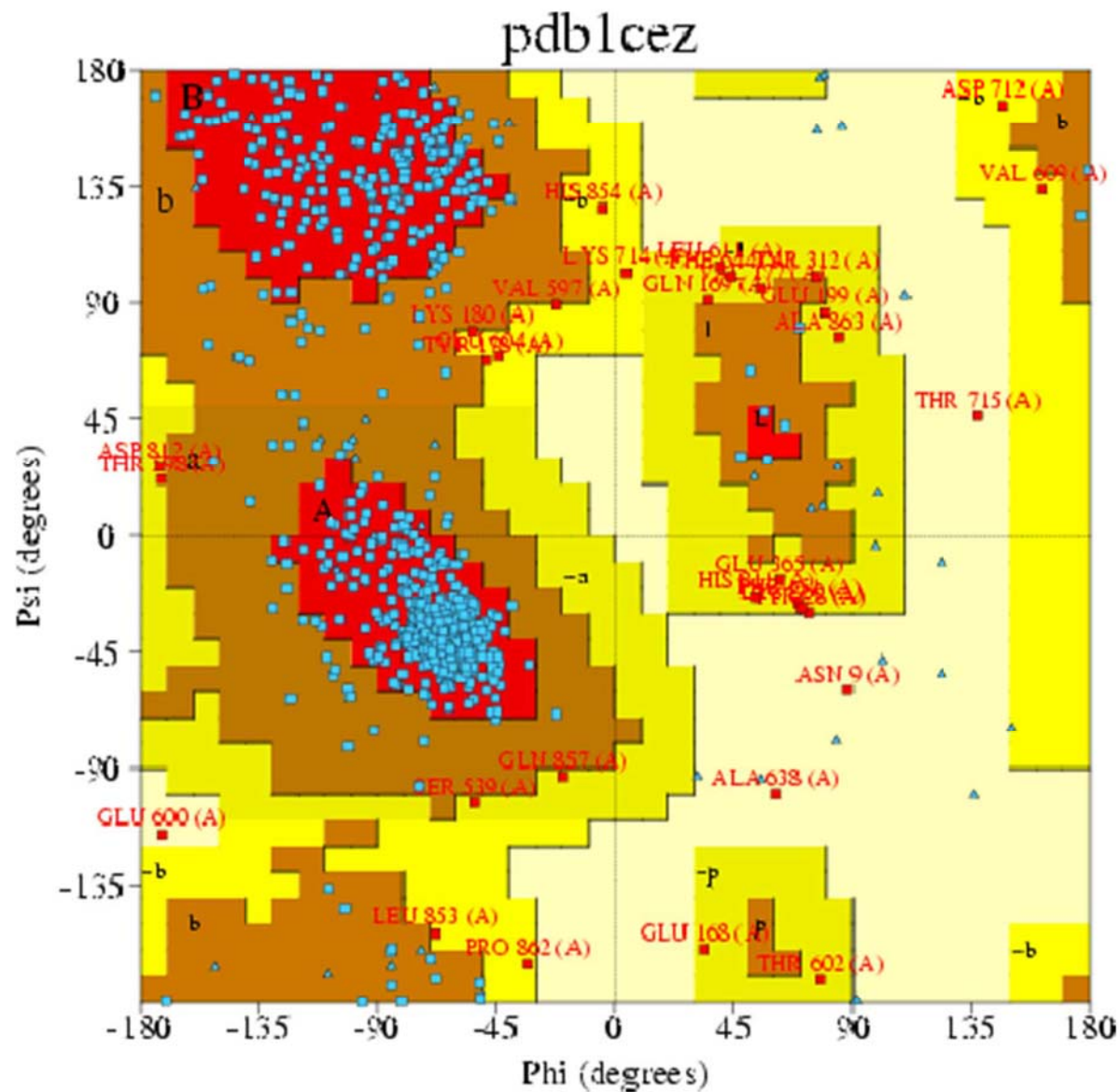
all helix



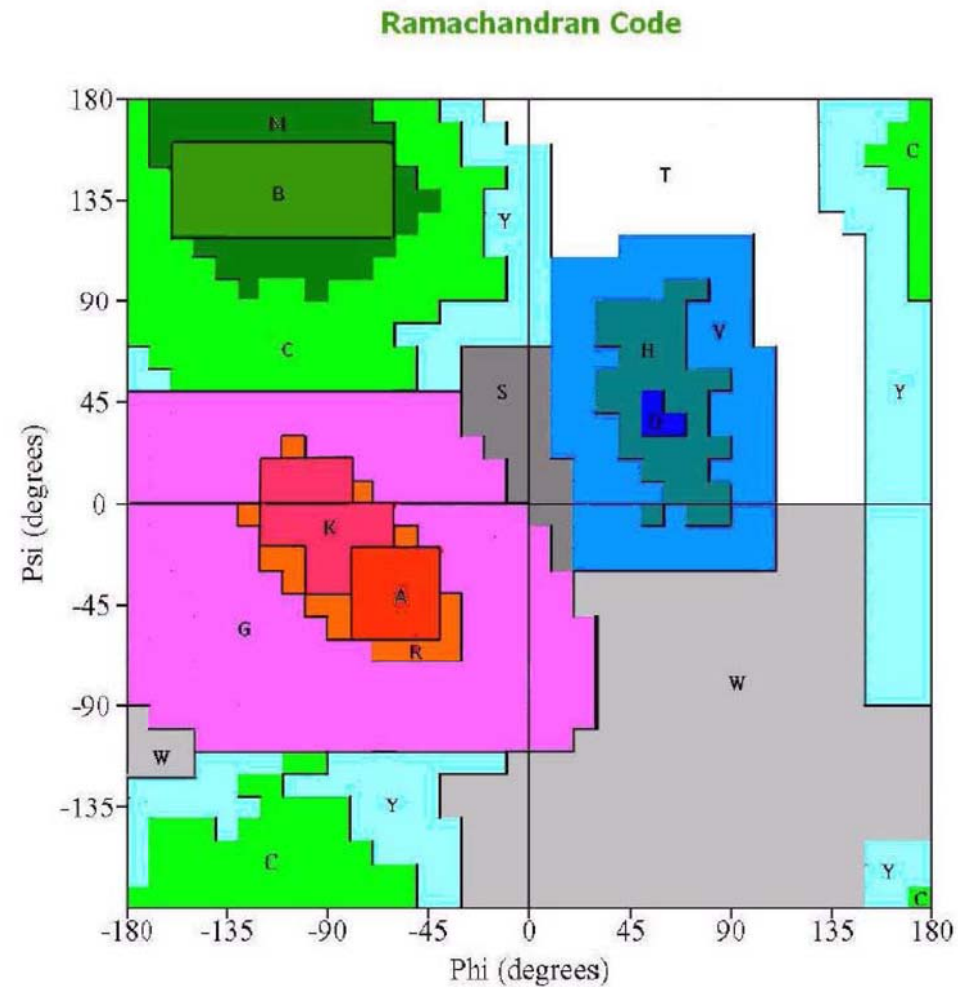
all beta



# The Ramachandran Map



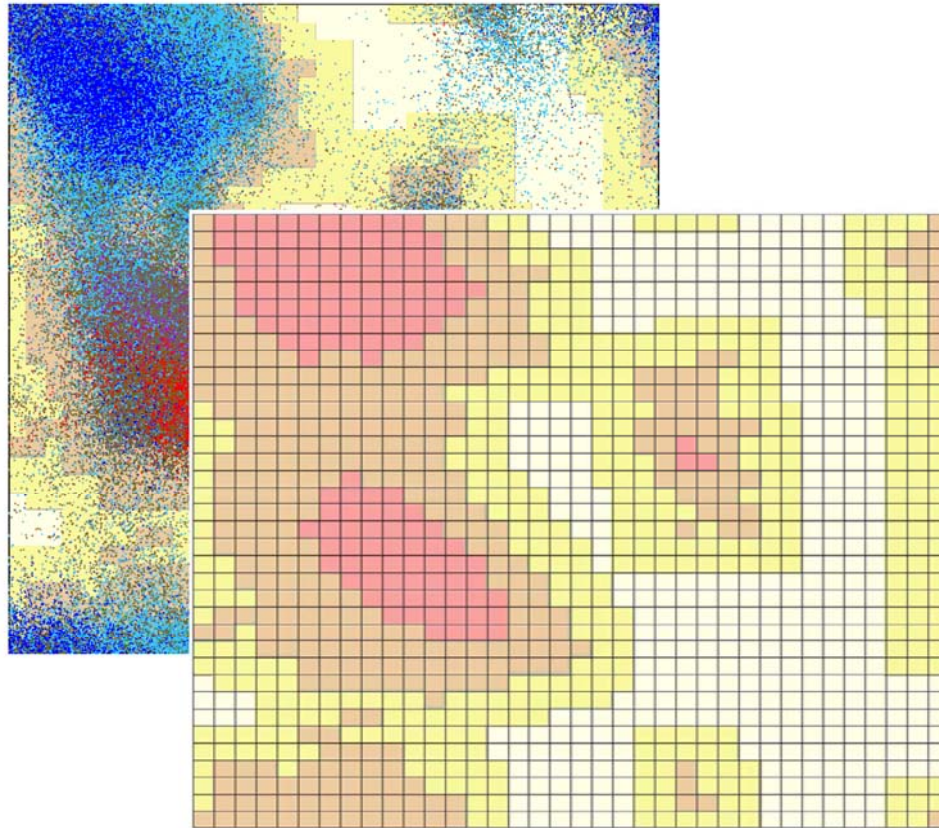
# An Example of Organized Ramachandran Map





SARST

## Distance Determination of the Cells

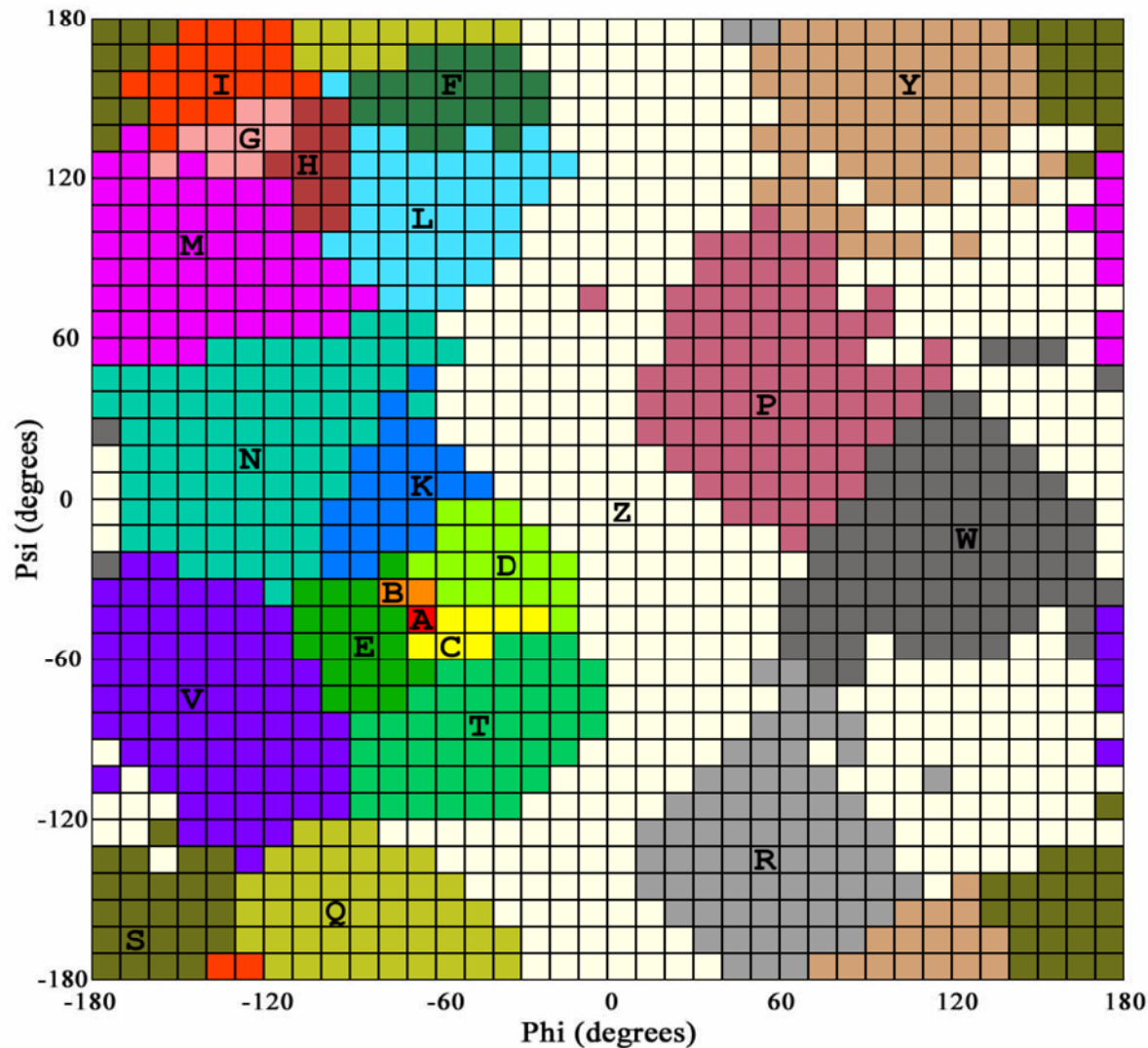


Root-Square Angular Distance

$$RSAD = \sqrt{(\Delta\phi)^2 + (\Delta\psi)^2}$$

# SARST

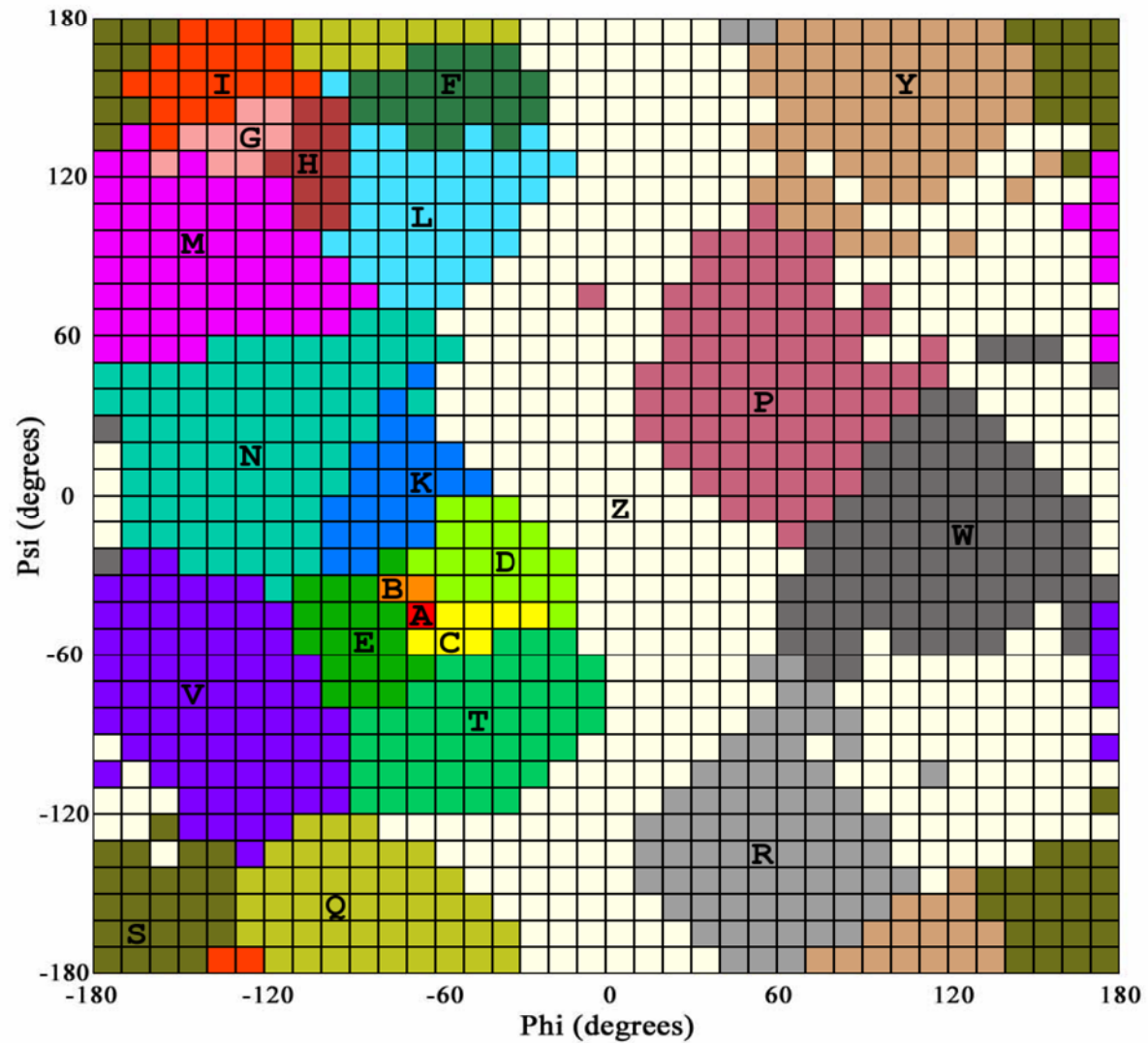
# Ramachandran Sequential Transformation



- Nearest-neighbor clustering
- 1,296 cells were clustered into 22 groups
- Each group was assigned with a symbol, i.e. Ramachandran code

SARST

# Ramachandran Sequential Transformation



RM Seq:  **I L L P E**

# How to Evaluate Similarities?

AAA<sup>W</sup>WW

AAAA<sup>W</sup>W

WWW<sup>A</sup>AA

WWWW<sup>A</sup>A

Are they equally similar?

Score A:A = ?

Score W:W = ?

Score A:W and W:A = ?

SARST

## BLOSUM-like Scoring Matrix for SARST

|   | A   | B   | C   | D  | E  | T  | K  | V  | N  | F  | G   | H   | I   | L  | M  | Q  | S   | Y   | R  | P  | W  | Z  | X |
|---|-----|-----|-----|----|----|----|----|----|----|----|-----|-----|-----|----|----|----|-----|-----|----|----|----|----|---|
| A | 3   | 2   | 2   | 1  | 1  | 0  | -2 | -3 | -3 | -8 | -11 | -11 | -13 | -8 | -8 | -9 | -14 | -9  | -7 | -8 | -7 | -4 | 0 |
| B | 2   | 2   | 2   | 1  | 1  | 1  | 0  | -1 | -2 | -6 | -12 | -10 | -10 | -7 | -7 | -6 | -10 | -8  | -5 | -6 | -4 | -6 | 0 |
| C | 2   | 2   | 2   | 1  | 1  | 3  | -1 | -2 | -3 | -6 | -13 | -11 | -9  | -7 | -8 | -7 | -9  | -10 | -2 | -7 | -5 | -3 | 0 |
| D | 1   | 1   | 1   | 3  | 1  | 2  | 2  | -1 | -1 | -4 | -9  | -7  | -8  | -4 | -6 | -5 | -7  | -4  | 1  | -3 | -4 | -2 | 0 |
| E | 1   | 1   | 1   | 1  | 3  | 1  | 2  | 3  | 1  | -5 | -7  | -6  | -7  | -4 | -4 | -4 | -7  | -2  | -1 | -5 | -3 | -1 | 0 |
| T | 0   | 1   | 3   | 2  | 1  | 5  | -1 | 2  | -1 | -2 | -6  | -6  | -4  | -4 | -5 | -2 | -4  | -4  | 2  | -1 | -1 | 3  | 0 |
| K | -2  | 0   | -1  | 2  | 2  | -1 | 4  | 1  | 3  | -3 | -6  | -6  | -5  | -3 | -3 | -2 | -5  | -2  | -2 | 0  | 0  | -1 | 0 |
| V | -3  | -1  | -2  | -1 | 3  | 2  | 1  | 9  | 3  | -3 | -4  | -4  | -2  | -2 | -2 | 0  | 0   | 3   | 2  | -1 | 3  | 4  | 0 |
| N | -3  | -2  | -3  | -1 | 1  | -1 | 3  | 3  | 5  | -2 | -4  | -4  | -3  | -2 | 0  | -2 | -3  | -2  | -1 | 1  | 1  | 1  | 0 |
| F | -8  | -6  | -6  | -4 | -5 | -2 | -3 | -3 | -2 | 5  | -1  | 1   | 0   | 3  | 0  | 3  | 0   | 2   | 0  | -2 | -2 | 1  | 0 |
| G | -11 | -12 | -13 | -9 | -7 | -6 | -6 | -4 | -4 | -1 | 4   | 3   | 3   | 0  | 2  | 0  | 1   | -3  | -5 | -5 | -6 | -2 | 0 |
| H | -11 | -10 | -11 | -7 | -6 | -6 | -6 | -4 | -4 | 1  | 3   | 4   | 1   | 2  | 2  | 0  | -1  | -2  | -4 | -3 | -5 | -1 | 0 |
| I | -13 | -10 | -9  | -8 | -7 | -4 | -5 | -2 | -3 | 0  | 3   | 1   | 4   | 0  | 1  | 2  | 4   | 0   | -1 | -4 | -7 | -2 | 0 |
| L | -8  | -7  | -7  | -4 | -4 | -4 | -3 | -2 | -2 | 3  | 0   | 2   | 0   | 4  | 1  | 1  | -1  | 0   | 0  | -1 | -2 | 1  | 0 |
| M | -8  | -7  | -8  | -6 | -4 | -5 | -3 | -2 | 0  | 0  | 2   | 2   | 1   | 1  | 4  | 0  | 1   | -1  | -4 | -2 | -2 | 1  | 0 |
| Q | -9  | -6  | -7  | -5 | -4 | -2 | -2 | 0  | -2 | 3  | 0   | 0   | 2   | 1  | 0  | 6  | 1   | 3   | 1  | -3 | -3 | 1  | 0 |
| S | -14 | -10 | -9  | -7 | -7 | -4 | -5 | 0  | -3 | 0  | 1   | -1  | 4   | -1 | 1  | 1  | 7   | 5   | 2  | -3 | -3 | 3  | 0 |
| Y | -9  | -8  | -10 | -4 | -2 | -4 | -2 | 3  | -2 | 2  | -3  | -2  | 0   | 0  | -1 | 3  | 5   | 10  | 7  | 2  | 2  | 7  | 0 |
| R | -7  | -5  | -2  | 1  | -1 | 2  | -2 | 2  | -1 | 0  | -5  | -4  | -1  | 0  | -4 | 1  | 2   | 7   | 11 | 3  | 0  | 7  | 0 |
| P | -8  | -6  | -7  | -3 | -5 | -1 | 0  | -1 | 1  | -2 | -5  | -3  | -4  | -1 | -2 | -3 | -3  | 2   | 3  | 8  | 7  | 4  | 0 |
| W | -7  | -4  | -5  | -4 | -3 | -1 | 0  | 3  | 1  | -2 | -6  | -5  | -7  | -2 | -2 | -3 | -3  | 2   | 0  | 7  | 9  | 5  | 0 |
| Z | -4  | -6  | -3  | -2 | -1 | 3  | -1 | 4  | 1  | 1  | -2  | -1  | -2  | 1  | 1  | 1  | 3   | 7   | 7  | 4  | 5  | 6  | 0 |
| X | 0   | 0   | 0   | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0   | 0   | 0   | 0  | 0  | 0  | 0   | 0   | 0  | 0  | 0  | 0  | 0 |

BLOSUM algorithm:

Henikoff and Henikoff. (1992) *Proc Natl Acad Sci USA*. **89**:10915-10919

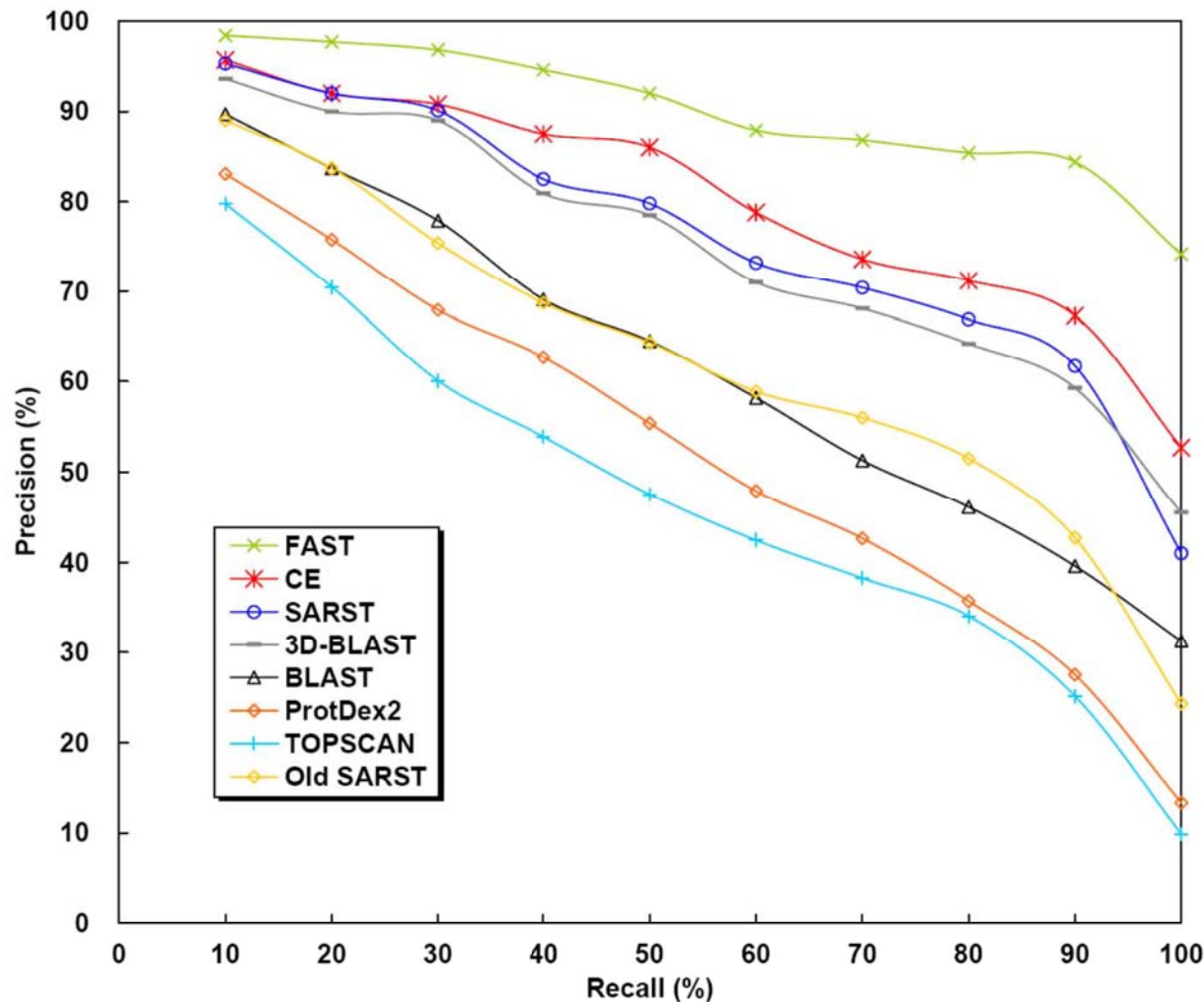
SARST

# Speed Evaluation

| Method         | Average time per query (sec) | Average time per comparison (sec) | Relative to SARST |
|----------------|------------------------------|-----------------------------------|-------------------|
| CE             | 82,789.20                    | 2.43E+00                          | 243,497.65        |
| FAST           | 6,241.57                     | 1.83E-01                          | 18,357.56         |
| TOPSCAN        | 85.08                        | 2.50E-03                          | 250.24            |
| YAKUSA         | 35.6                         | 1.05E-03                          | 104.71            |
| 3D-BLAST       | 9.07                         | 2.66E-04                          | 26.68             |
| ProtDex2       | 0.76                         | 2.23E-05                          | 2.24              |
| BLAST          | 0.30                         | 8.76E-06                          | 0.88              |
| SARST          | 0.34                         | 9.98E-06                          | 1.00              |
| SARST (2 CPUs) | 0.16                         | 4.70E-06                          | 0.47              |



# Accuracy Evaluation



## Information retrieval

- **Recall**  
the ability to extract answers
- **Precision**  
the ability to give correct answers

***Next...***



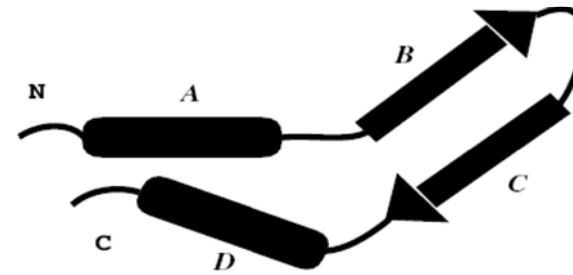
<http://sarst.life.nthu.edu.tw/iSARST>

# **CPSARST - Circular Permutation Search Aided by Ramachandran Sequential Transformation**

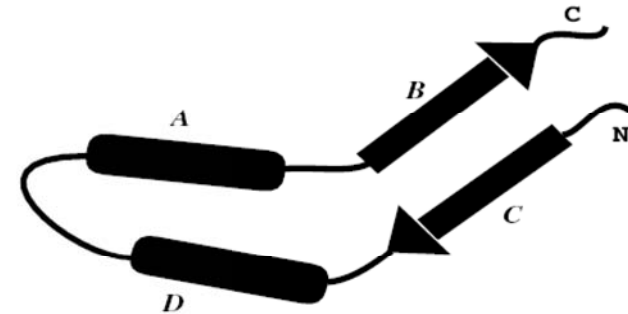
Lo WC, Lyu PC: ***CPSARST: an efficient circular permutation search tool  
applied to the detection of novel protein structural relationships.***  
Genome Biology 2008,9:R11.

# Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere<sup>1</sup>.
- In most of the cases, naturally occurring CPs have similar 3D structures and conserved biological functions<sup>2</sup>.
- Efficient CP search tool is not available yet.



*The sequence: ..A..B..C..D..*



*The sequence ..C..D..A..B..*

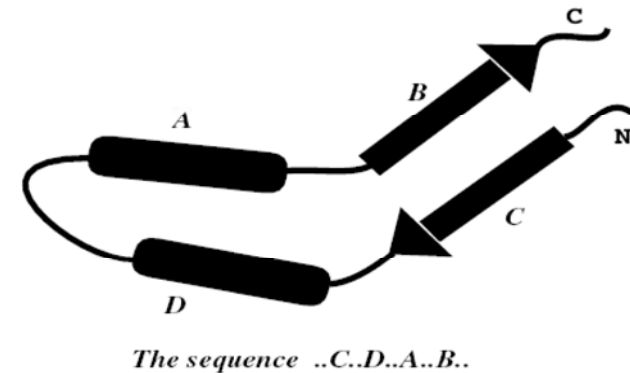
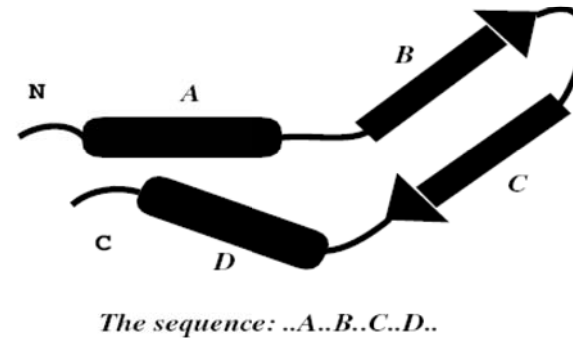
1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.

# Natural Circular Permutants

- Plant lectins
- Transaldolases
- DNA and other methyltransferases
- Ferredoxins
- Proteinase inhibitors
- Bacterial  $\beta$ -glucanases
- Swaposins
- Glucosyltransferases
- $\beta$ -glucosidases
- SLH domains
- C2 domains
- FMN-binding proteins
- Double- $\phi\beta$ -barrels
- Glutathione synthetases

# Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere<sup>1</sup>.
- In most of the cases, CPs have similar 3D structures and conserved biological functions<sup>2</sup>.
- Efficient CP search tool is not available yet.



1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.



# Applications of Circular Permutation

- Folding researches.
- Determination of structurally and functionally important segments<sup>1,2</sup>.
- Modification (enhancement) of the activity and/or stability<sup>3-5</sup>.
- Creation of novel fusion proteins, the tethered sites of which are not confined to the native termini<sup>5,6</sup>.

1. Anand.B. et al. Nucleic Acid Res 2006;34:2196-2205.

2. Gebhard.LG. et al. J Mol Biol 2006;358:280-288.

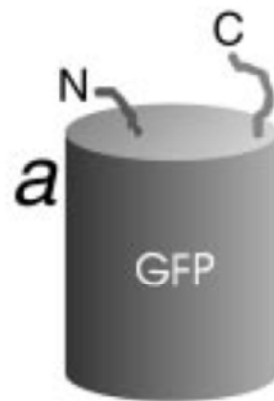
3. Qian.Z., Lutz.S. J Am Chem Soc 2005;127:13466-13467.

4. Schwartz.TU. et al. Protein Sc 2004;13:2814-2818.

5. Kojima.M. et al. J Biosci Bioeng 2005;100:197-202

6. Baird.GS. et al. Proc Natl Acad Sci USA 1999;96:11241-11246.

# Fluorescent Calcium Sensor with CP

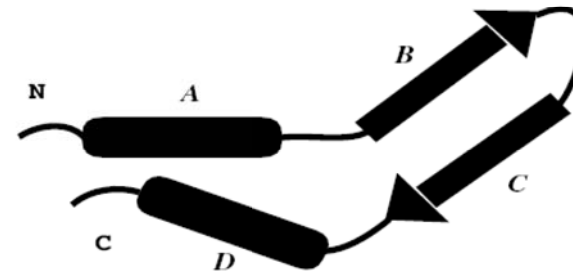


G.S. Baird, et al. **Circular permutation and receptor insertion within green fluorescent proteins.** *PNAS* 1999;96:11241-11246

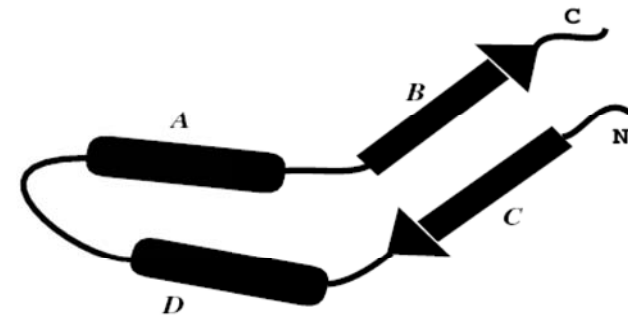


# Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere<sup>1</sup>.
- In most of the cases, naturally occurring CPs have similar 3D structures and conserved biological functions<sup>2</sup>.
- **Efficient CP search tool is not available yet.**



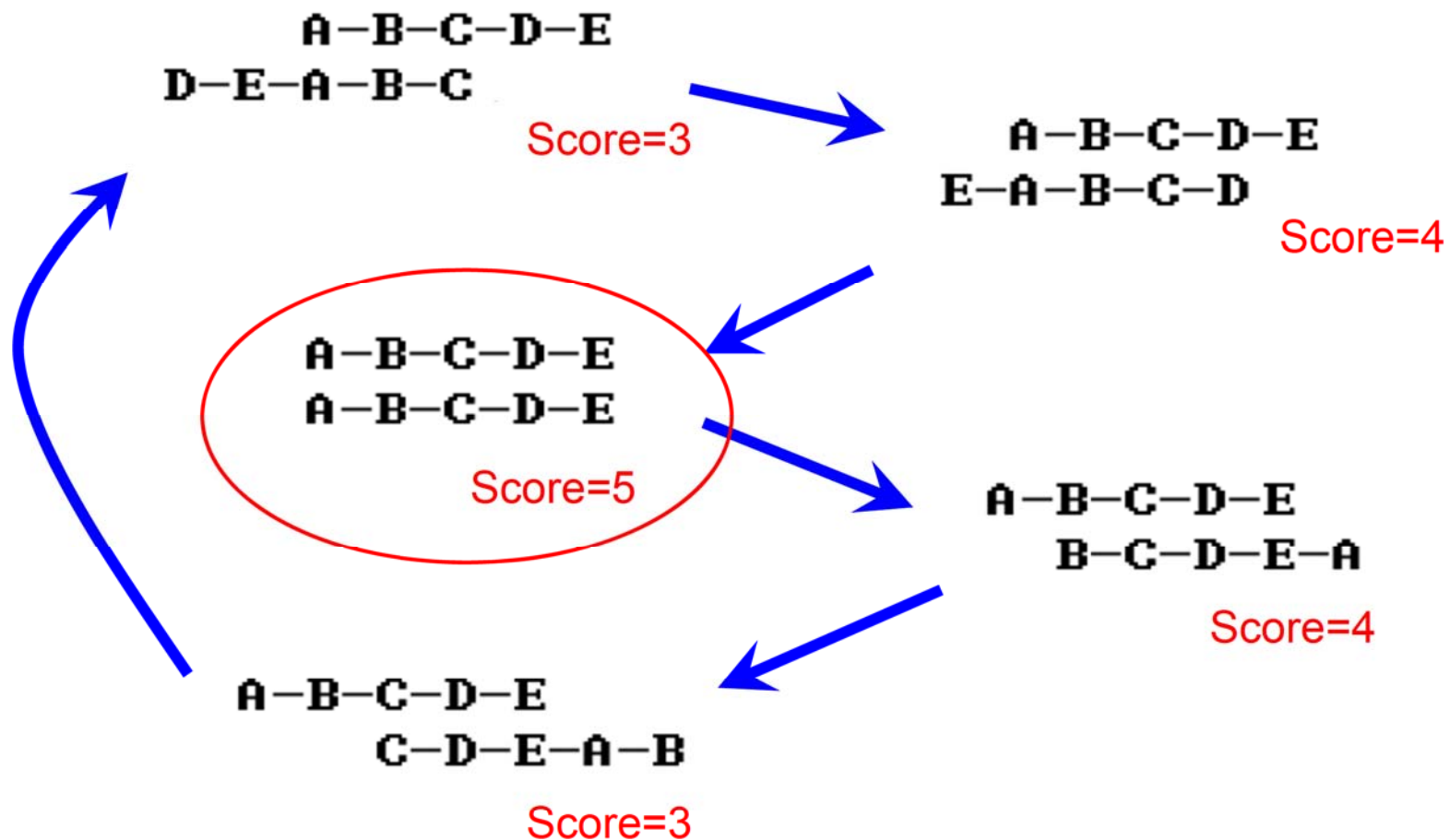
*The sequence: ..A..B..C..D..*



*The sequence ..C..D..A..B..*

1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.

# Basic Approach to the Detection of CP

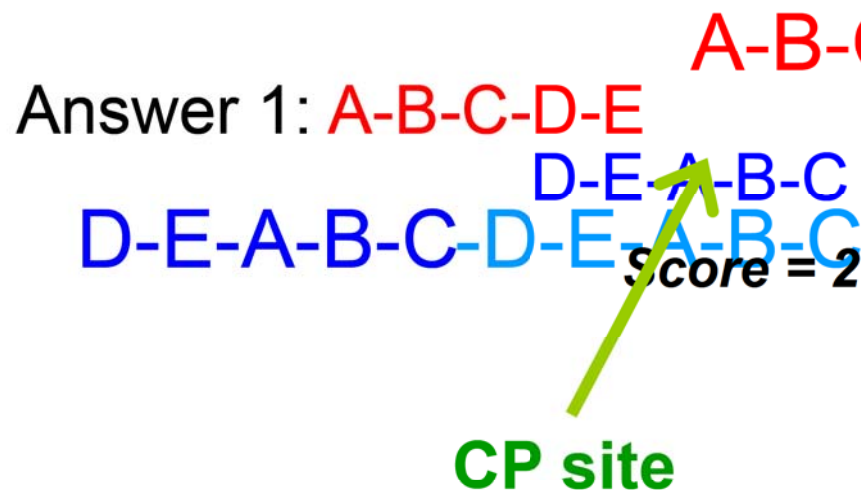


# The Basic Idea of CPSARST

Target: A-B-C-D-E

Query: D-E-A-B-C

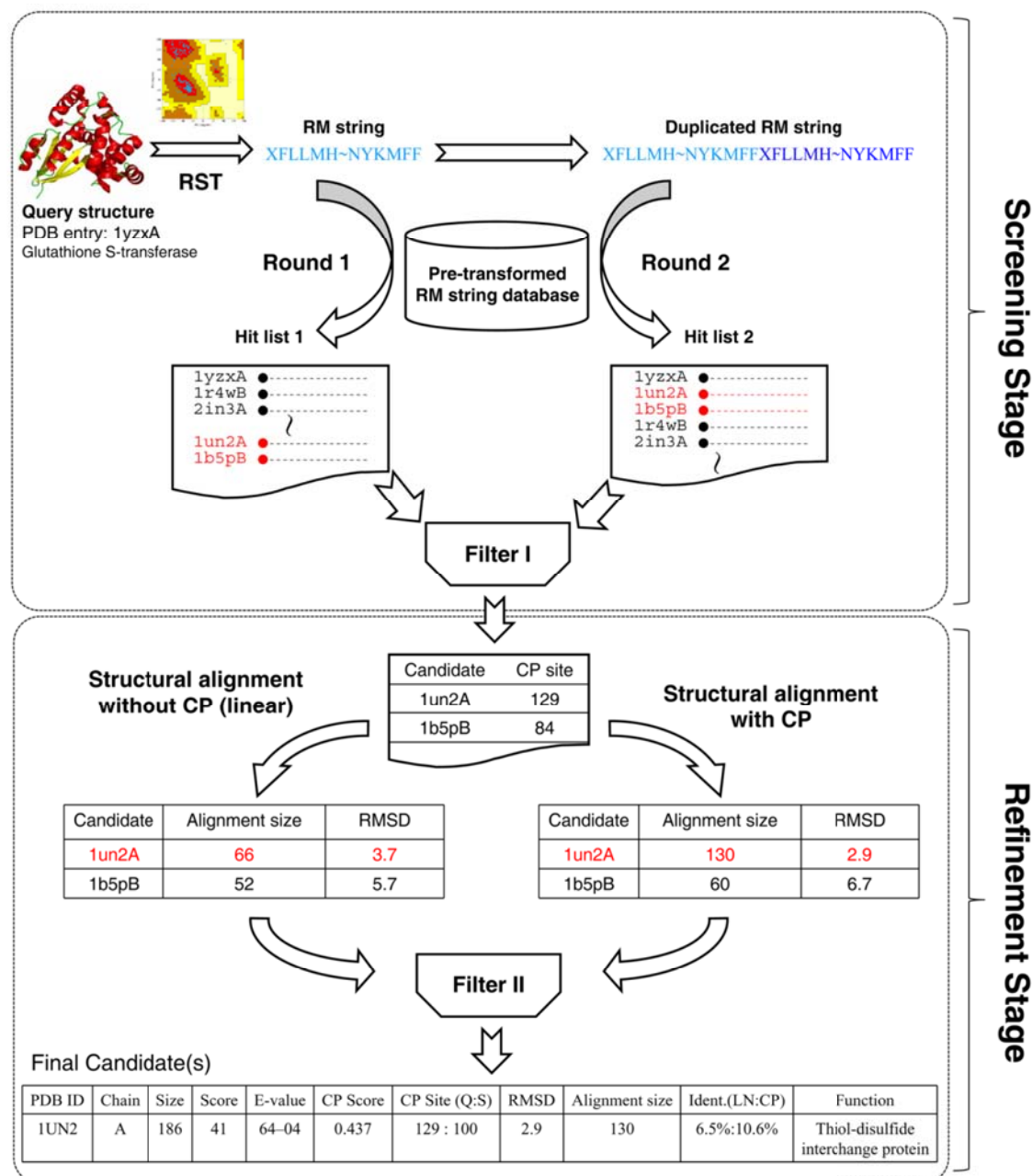
Answer 1: A-B-C-D-E  
D-E-A-B-C  
D-E-A-B-C-D-E-A-B-C  
Score = 2  
CP site



Answer 2: A-B-C-D-E  
D-E-A-B-C  
Score = 5  
Score = 3



# The Double Filter-and-Refine Strategy





## Statistics of protein structural database searches by CPSARST

| Database                                |                                         |              | nrPDB-90            | nrSCOP-90           |
|-----------------------------------------|-----------------------------------------|--------------|---------------------|---------------------|
| No. of proteins                         |                                         |              | 14,422              | 11,688              |
| No. of<br>candidate<br>pairs            | 1. Detected by amino acid sequence      |              | 5,020               | 1,802               |
|                                         | 2. Detected only by Ramachandran string |              | 252,287             | 196,533             |
|                                         | 3. Confirmed after the refinement stage | Total        | 2,911               | 4,228               |
|                                         |                                         | Symmetric CP | 682                 | 1,161               |
| Total No. of protein pairs              |                                         |              | $208.0 \times 10^6$ | $136.6 \times 10^6$ |
| Total running time (minutes)            |                                         |              | 3,942               | 1,974               |
| No. of protein pairs scanned per minute |                                         |              | 52,764              | 69,204              |

# Speed Advantage of CPSARST

- 4 times faster than [UFAU](#) (sequence-based)
  - Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999, **15**:930-936.
- 8,824 times faster than SAMO (structure-based)
  - Chen L et al.: **Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison.** *BMC Struct Biol* 2006, **6**:18.
- CPSARST requires only 1.7 minute to scan the current PDB (~90,000 polypeptides).

# Performance of pair-wise comparisons for natural candidate CP pairs over various sequence identities

$$\left( \frac{\text{Alignment size}}{\text{Average protein size}} \right)^{1.5}$$

| Identity (%) | No. of candidate CP pairs | Structural diversity |        |              |
|--------------|---------------------------|----------------------|--------|--------------|
|              |                           | CPSARST              | SHEBA  | SAMO         |
| ≤ 10         | 823                       | <u>6.309</u>         | 11.180 | <u>4.396</u> |
| 10 ~ 20      | 152                       | <u>5.864</u>         | 13.881 | <u>4.994</u> |
| 20 ~ 30      | 11                        | 3.581                | 4.506  | 3.363        |
| 30 ~ 40      | 33                        | 1.868                | 3.284  | 2.210        |
| 40 ~ 50      | 40                        | 1.755                | 3.096  | 1.544        |
| > 50         | 9                         | 1.385                | 2.247  | 1.520        |

## Top 20 homologs retrieved from nrPDB by DALI for hypothetical protein YlqF

| No. | PDB entry / Size | Function                                   |
|-----|------------------|--------------------------------------------|
| 1   | 1pujA / 261      | Conserved hypothetical protein YlqF        |
| 2   | 1u0lA / 278      | Probable GTPase                            |
| 3   | 1ctqA / 166      | p21h-Ras-1 fragment                        |
| 4   | 1ejjA / 508      | Phosphoglycerate mutase (isomerase)        |
| 5   | 1gpmA / 501      | Amidotransferase, GMP synthetase           |
| 6   | 1efcA / 386      | Elongation factor Eftu (RNA binding)       |
| 7   | 1hrkA / 359      | Ferrochelatase fragment (lyase)            |
| 8   | 1ni5A / 428      | Putative cell cycle protein Mesj           |
| 9   | 1dpgA / 485      | Glucose 6-phosphate reductase              |
| 10  | 2hjaA / 390      | GTP-binding protein engA                   |
| 11  | 1veeA / 134      | Unknown function proline-rich protein      |
| 12  | 1cqxA / 403      | Flavohepotein (lipid binding)              |
| 13  | 2p8zT / 813      | Elongation factor 2                        |
| 14  | 1mkyA / 400      | Probable GTP-binding protein               |
| 15  | 1dar / 615       | Elongation factor G (translational GTPase) |
| 16  | 1kk1A / 397      | Eif2gamma mutant                           |
| 17  | 1hurA / 180      | Human ADP-ribosylation factor 1            |
| 18  | 1fdr / 244       | Flavodoxin reductase                       |
| 19  | 2clsA / 179      | Rho-related GTP-binding protein            |
| 20  | 1wcwA / 254      | Uroporphyrinogen III synthase              |
| 21  | 1ak1 / 308       | Ferrochelatase                             |

Top 20 circular permutants detected from nrPDB by CPSARST for  
hypothetical protein YlqF

| No. | PDB entry / Size | Function                                  |
|-----|------------------|-------------------------------------------|
| 1   | 1ZBD / 203       | Rabphilin-3A                              |
| 2   | 1KY2 / 182       | GTP-binding                               |
| 3   | 2F7S / 217       | Ras-related protein Rab-27B protein YPT7P |
| 4   | 2NZJ / 175       | GTP-binding protein REM 1                 |
| 5   | 1T91 / 207       | Ras-related protein Rab-7                 |
| 6   | 1X3S / 195       | Ras-related protein Rab-18                |
| 7   | 1YU9 / 175       | GTP-binding protein, GTPase domain        |
| 8   | 2EW1 / 201       | Ras-related protein Rab-30                |
| 9   | 2GF9 / 189       | Ras-related protein Rab-3D                |
| 10  | 1YVD / 169       | Ras-related protein Rab-22A               |
| 11  | 1PUI / 210       | Probable GTP-binding protein engB         |
| 12  | 2O52 / 200       | Ras-related protein Rab-4B                |
| 13  | 1U8Y / 168       | Ras-related protein Ral-A                 |
| 14  | 1HUQ / 164       | Rab5C, GTPase domain                      |
| 15  | 2HUP / 201       | Ras-related protein Rab-43                |
| 16  | 1FZQ / 181       | ADP-ribosylation factor-like protein 3    |
| 17  | 2OCB / 180       | Ras-related protein Rab-9B                |
| 18  | 1OIV / 191       | Ras-related protein Rab-11A               |
| 19  | 2FN4 / 181       | Ras-related protein R-Ras                 |
| 20  | 1Z0F / 179       | Rab14, member Ras oncogene family         |



# Multiple Alignment of Raw Sequences

```

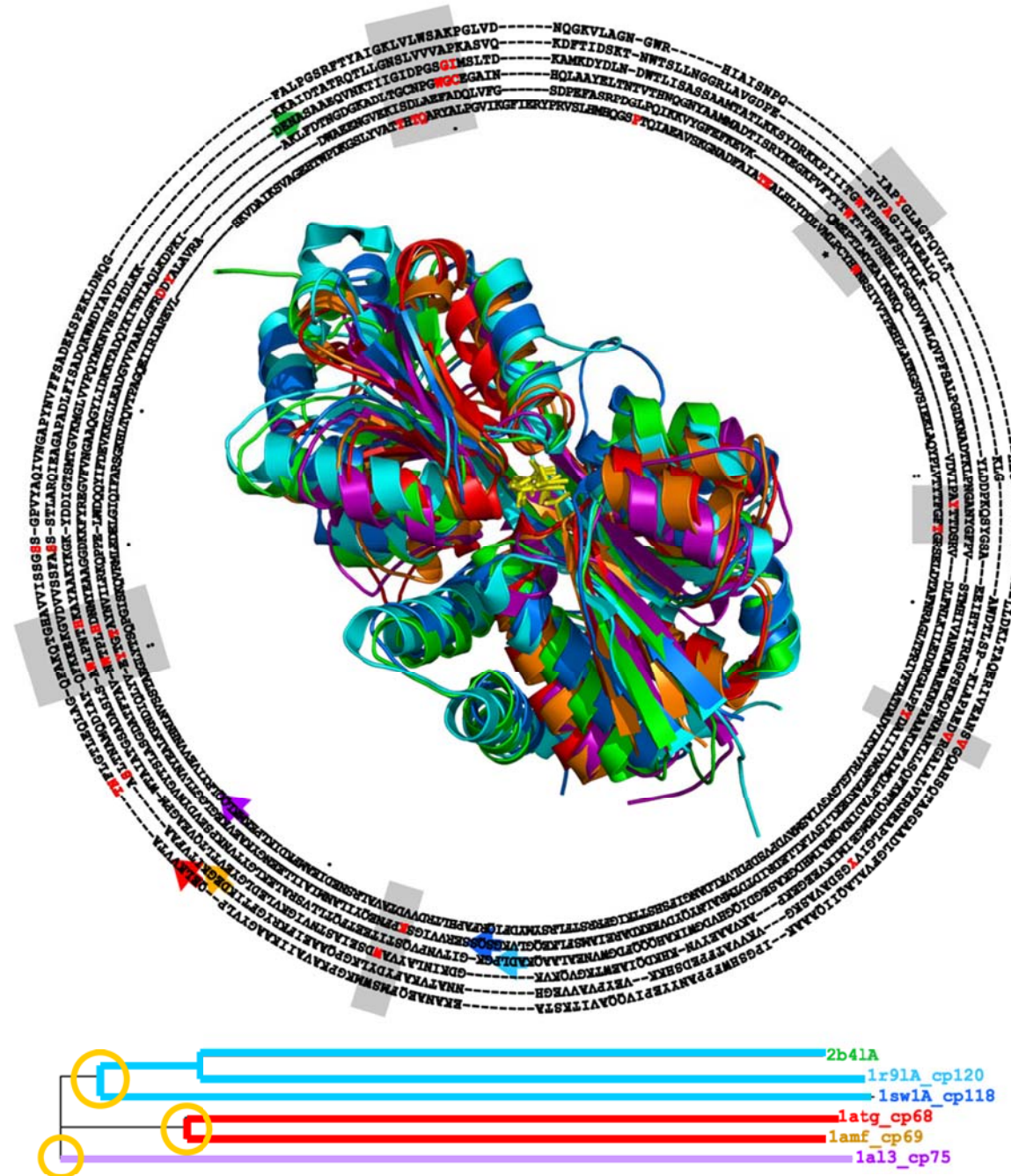
latg  -----ELKVVTATNFLGTLEQLAGQFAKQTHAVVISSGSSSGPVYAQIVNGAPYNVFFSADEKSPEKLDN-----QGFALPG 72
lamf  -----DEGKITVFAAASLTNAMQDIATQFKKEGVDVSSFASSSTLARQIEAGAPADLFISADQKWMDYAVD-----KKAIDTA 75
lal3  MKLQQLRYIVEVNVNHNLSSTAEGLYTSQPGISKQVRMLEDELGIQIFARSGKHLTQVTPAGQEIIIRIAREVLSKVDAIKSVAGEHTWPDKGSLYVATHTQARYALPGV-IKGFIERY 119
lsw1A -----GSQSSERVVIGSKPFNEQYILANMIAILLEENGYKAEVKEGLGGTLVNYEALKRNDIQLYVEYTGTAYNVILRKQPPELWDQQYIFDEVKKGLLEADG--VVVAAKLG 106
2b41A -----DENASAAEQVNKTIIGIDPGSGIMSLTDKAMKDYDLNDWTLISASSAAMTATLKSYDRKKPIIITGWTPHWMFSRYKLKYLDDPKQSYGSAEEIHTI 98
1r91A -----ADLPKGGITVNPVQSTITEETFQTLVSRALEKLGYTVNKPSEVDYNVGYTSLASGDATFTAVNWTPLHDNMYEAAGGDKKFYREGVFVNGAAQGYLIDKKTADQ 105
.
latg  SRFTYAIGKLVLSAKPGLVDNQGKVLAGNGWR-----HIAISNPQIAPYGLAGTQVLTHLGLLD-----KLTAQERIVEANSVGQAHSQTASGA 157
lamf  TRQTLLGNSLVVAPKASVQKDFT-IDSKINWTSLLN-----GGRLAVGDPEHVPAGIYAKEALQKLGAWD-----TLSP--KLAPAEDVRGALALVERNE 163
lal3  PRVSLHMHQGSPTQIAEAVSKGNADFAIATEALHLYDDLVMLPCYHWNRSIVVTPEHPLATKGSVSIEELAQYP-----LVITYTFGTGRSELDTAFNRAGLTP 218
lsw1A FRDDYALAVRADWAEENGVEKISDLAEFADQLVFGSD-----PEFASRPDGLPQIKKVYGFEFKEVKQME-----PTLMEAIKKNQVDVIPAYTTDSRV 196
2b41A TRKGSKEQPNAKLLSQFKWTQDEMGEIMIKVEEGE-----KPAKVAAEYVNKHKDQIAEWTKGVQKVK-----GDKINLAYVAWDSEIASTNVIGKVL 188
1r91A YKITN-IAQLKDPKIAKLFDTNGDGKADLTGCNPGWGCEGAINHQLAAYELTNTVIHNQGNYAAMMADTISRYKEGKPVFYYTWIPYWSNELKPGKDVWLQVPFSALPGDKNADTKLP 224
:
latg  ADLGFVALAQIIQAAAKIPGSHWFPANYEPIVQQAVITKST-----AEKANAEQFMSWMK--GPKAVAIIKAAGYVLPQ-----231
lamf  APLGIVYGSDAVASKG-VKVVATFPEDSHKK--VEYPVAVVEG-----HNNATVKAFYDYLK--GPQAAEIFKRYGFTIK-----233
lal3  RIVFTATDADVIKTYVRLGLGVGVIASMAVDPVSDPDLVKLDANGIFSHSTTKIGFRRSTFLRSYMYDFIQRFAPHLTRDVVDTAVALRSNEDIEAMFKDIKLPEK 324
lsw1A DLFNLKILEDDKGALPPYDAIIIVNGNTAKDEKLISVLKLEDR-----IDTDTMRALNYQYDVEKKDAREIAMSFLKEQGLVK-----275
2b41A EDLGYEVTLTQVEAGPMWTAIATGSADASLSAWLPNTHKAYAAKYKG-----KYDDIGTSMGTGVKMGLVPPQYMKNVNSIEDLKK-----268
1r91A NGANYGFPVSTMHIVANKAWAEKNPAAAKLFAIMQLPVADINAQNAIMHDG---KASEGDIQGHVDGWIKAHQQQFDGWVNEALAAQK-----309

```





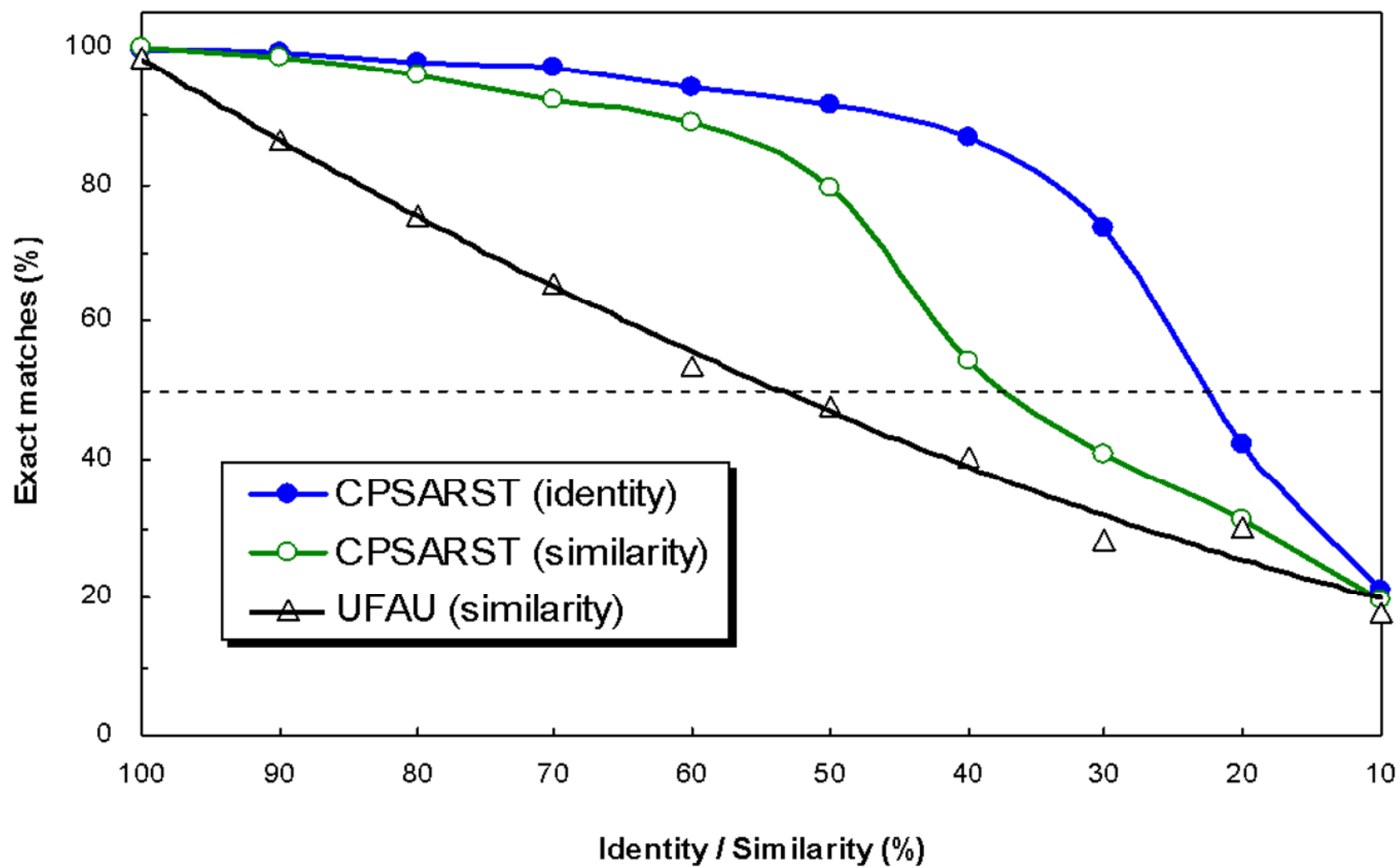
# Multiple Alignment of Circularly-Permuted Sequences



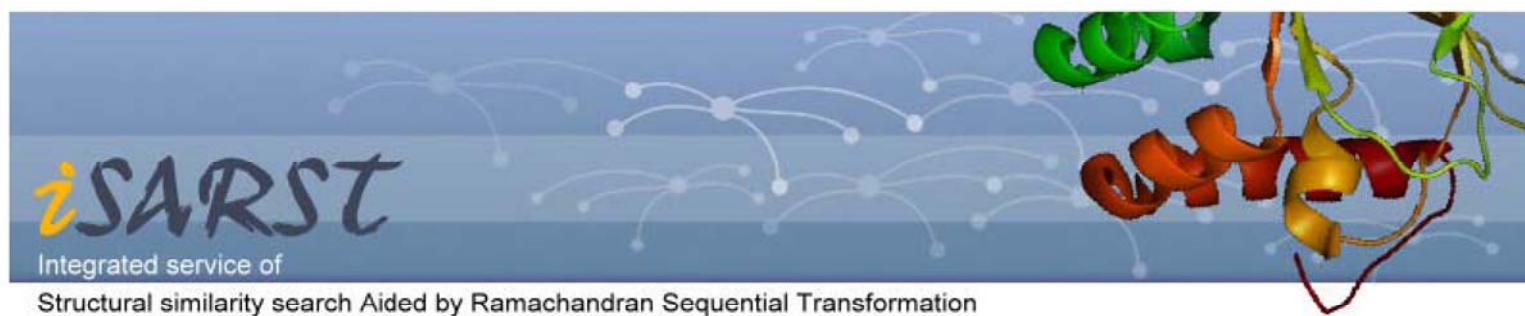
## Possible Applications of CPSARST

- Bank-against-bank searches are achievable.
- Develop automated procedures such as the functional assignment system for novel hypothetical proteins
- Construct CP database

(a)



***Next...***



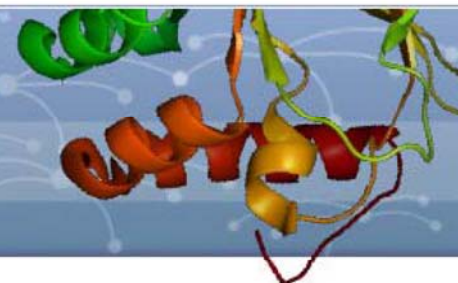
<http://sarst.life.nthu.edu.tw/iSARST>

<http://sarst.life.nthu.edu.tw/iSARST/>

**iSARST**

Integrated service of

Structural similarity search Aided by Ramachandran Sequential Transformation



## Welcome to iSARST

 [Tutorial](#)

Currently 63 threads are running on this PC-cluster.

A typical search along with superimposing 100 structures takes only **3~5 seconds**.

Circular permutants can be identified, even when the sequence identity is **<10%** (*Example [pair](#) / [family](#)*).

**Please enjoy the speed, accuracy and convenience brought about by iSARST!**

Query

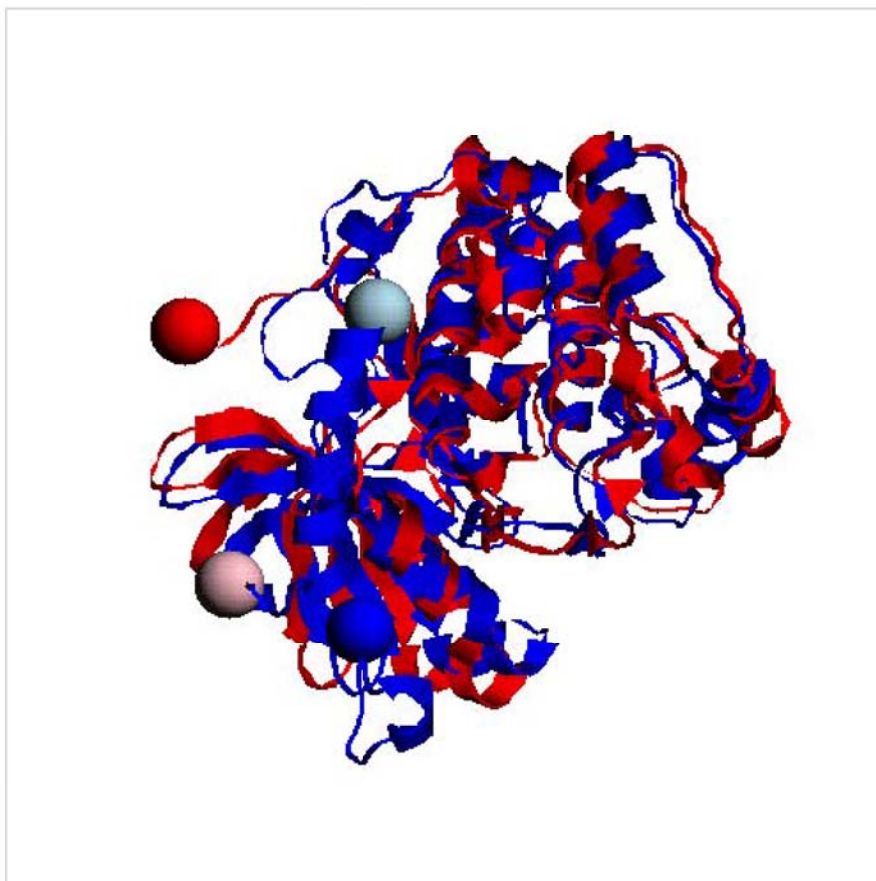
- ☐ 5-letter PDB or 7-letter PDB entry(s):   
\* Format: PDB id + chain ID. If there is no chain ID, simply use the 4-letter PDB id or use '\_' to represent the chain.  
Multiple entries are acceptable (batch mode), please use the comma to separate them.  
Example: 1atpE, 1cewI, 1ti5A, 1JUL, 1HEL\_, d1swya\_, d1oxda\_
- ☐ Local PDB file:    
& Chain ID in this file:   
\* If there is no chain ID, please leave it blank or use '\_' to represent it.  
You can also use '\*' as the chain ID and then every chain will be used to search the database.
- ☐ Compressed PDB collection:    
File type: ☐ .zip ☐ .rar ☐ .tar.gz ☐ .tar  
\* To perform SARST in this batch mode, please specify a compressed file collecting PDB structures,  
choose the correct file type, and then click "Submit" (Maximum size = 16M).
- ☐ Previous session ID:  [Select a previous session...](#)   
\* Previous searching results can be retrieved by using session ID.

[Sample: co-linear search](#)

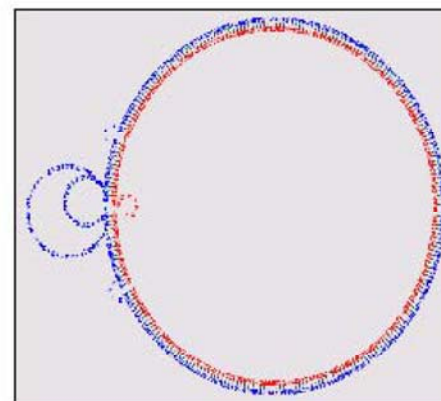
[Sample: circularly-permuted search](#)



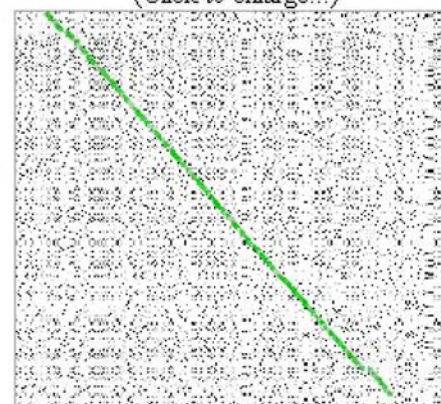
## Structural Alignment by FAST



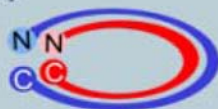
|                      |                                  |
|----------------------|----------------------------------|
| <b>Qry</b>           | <a href="#">1atpE</a> (336 a.a.) |
| <b>Sbj</b>           | <a href="#">2vgpA</a> (264 a.a.) |
| Aligned residues     | 254 a.a.                         |
| RMSD                 | 2.421                            |
| Structural diversity | 3.108                            |
| Identity             | 29.2% (77/264)                   |
| Similarity           | 50.4% (133/264)                  |



(Click to enlarge...)



Key


☒ Qry  
only  
☒  
backbone

☒ Sbj only  
☒ reset

Close this window

To view the superimposed structures,  
you may need MDL® Chime freely available at  
<http://www.mdl.com/products/framework/chime/>



# Tutorial of *i*SARST

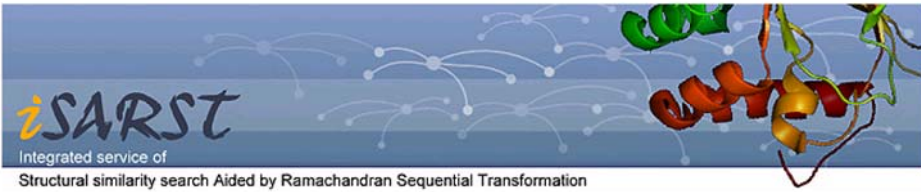
[Home](#) | [Downloads](#) | [References](#) | [Tutorial](#)



Tutorial

- How to make a request?
  - [Input PDB entry/entries](#)
  - [Upload a PDB file](#)
  - [Upload an archive file](#)
  - [Retrieve previous results](#)
- [Settings](#)
- [Final output](#)
- [How to install Chime in Firefox?](#)

[Browse the sample results](#)



**iSARST**  
Integrated service of  
Structural similarity search Aided by Ramachandran Sequential Transformation

☐ 5-letter PDB entry(s):

\* Format: PDB id + chain ID. If there is no chain ID, simply use the 4-letter PDB id or use '\_' to represent the chain. Multiple entries are acceptable (batch mode), please use the comma to separate them. Example: 1a0E, 1cew, 1b5A, 1JUL, 1HEL\_

☐ Local PDB file:

& Chain ID in this file:

\* If there is no chain ID, please leave it blank or use '\_' to represent it. You can also use '\*' as the chain ID and then every chain will be used to search the database.

☐ Compressed PDB collection:

File Type: ☐ .zip ☐ .rar ☐ .tar.gz ☐ .tar

\* To perform SARST in this batch mode, please specify a compressed file collecting PDB structures, choose the correct file type, and then click "Submit" (Maximum size = 1639).

☐ Previous session ID:

\* Previous searching results can be retrieved by using session ID.

|                   |                                                                                                                                                                                                                                                             |
|-------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Subject type      | <input checked="" type="radio"/> Co-linear structural homologs (search engine: SARST)<br><input type="radio"/> Circularly-permuted structural homologs (search engine: CPSARST)                                                                             |
| Target database   | 100% identity non-redundant PDB (Aug. 2006) 45,336 polypeptides                                                                                                                                                                                             |
| Parameters        | Hit list size: <input type="text" value="100"/> E-value cutoff: <input type="text" value="1e-7"/><br>Gap-opening (G) and Gap-extension (E) Penalties: <input type="text" value="G=9, E=2 (Highest Accuracy)"/> <input type="button" value="Load defaults"/> |
| Refinement engine | <input checked="" type="radio"/> FAST (Zhu and Weng, 2005)<br><input type="radio"/> TM-align (Zhang and Skolnick, 2005)                                                                                                                                     |

## Welcome to *i*SARST

In this service, we implement two protein structural similarity search methods, [SARST](#) and [CPSARST](#). Besides, outstanding structural alignment tools, [FAST](#), [TM-align](#) and [SAMO](#), are recruited as the refinement engines. The state-of-the-art algorithm for improving the quality of structure-based sequence alignment [SE](#) is also implemented here. We would like to thank these authors for their excellent developments, which have greatly moved this research field forward.

*i*SARST allows users to input many structures at once. Its MPI system will do the similarity searches and structural alignments in a batch mode to rapidly

<http://sarst.life.nthu.edu.tw/iSARST/hlp/tutorial.php>

54

# CPDB:

## a database of circular permutation in proteins



### CPDB - the Circular Permutation Database

#### Welcome to the CPDB

Circular permutation (CP) of a protein can be visualized as if its original termini were linked and new ones created elsewhere. Since the first observation of CP in plant lectins, a substantial number of natural examples have been reported, including  $\beta$ -glucanases, swaposins, glucosyltransferases,  $\beta$ -glucosidases, SLH domains, transaldolases, C2 domains, FMN-binding proteins, double- $\alpha$   $\beta$ -barrels, glutathione synthetases, methyltransferases, ferredoxins, and proteinase inhibitors. In most of the cases, circular permutants (CPs) have conserved function or enzymatic activity, sometimes with increased functional diversity.

To reveal the influences of CP on the structure, function and folding of proteins, many artificial CPs have been generated, such as trypsin inhibitor, anthranilate isomerase, dihydrofolate reductase, T4 lysozyme, ribonucleases, aspartate transcarbamoylase,  $\alpha$ -spectrin SH3 domain, DsbA protein, ribosomal protein S6 and  $\beta$ -glucanase. The outcomes have indicated that protein structures seem remarkably insensitive to CP and, CPs generally retain their biological functions with sometimes increased stability or activity. Because of this, CP has been applied to trigger crystallization, improve enzyme activities, determine critical elements, and create novel fusion proteins, the tethered sites of which are not confined to the native termini. Recently, it has also been reported that the CP relationship among proteins can be used to assign possible functions for novel hypothetical proteins (see [CPSARST](#)). However, in spite of these interesting properties and applications, there is still much uncertainty about the genetic mechanisms, the evolutionary importance and the natural prevalence of CP.

The CPDB provides resources for studying CP and CP relationships among protein structures. This site also offers viable CP site predictions in order to facilitate the application of CP in academic researches and biotechnological developments.

#### Methods

Primary data of CPDB were collected from the non-redundant PDB dataset by using [CPSARST](#). [FASE](#) and visual inspections were then performed to refine the data. Methods described by [Paszkiwicz KH et al.](#) are implemented to predict other viable CP sites for the circular permutants identified. [FAST](#) is recruited in the website as the structural alignment engine.

#### Statistics

The non-redundant subset of CPDB contains about 11%, 32% and 57% mainly- $\alpha$ , mainly- $\beta$  and  $\alpha$ - $\beta$  mixed protein structures, respectively.

[Browse the Hierarchy](#)

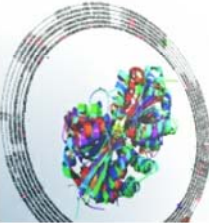
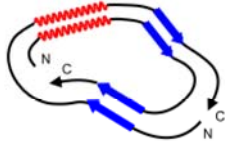
[Batch Browsing](#)

[Circular Permutation Search](#)

[Structural Similarity Search](#)

[CP-related Publications](#)

  
  
Search by  
☐ PDB ID  
☒ Keyword



<http://sarst.life.nthu.edu.tw/cpdb/>

**Thanks for your attentions.**

